Encyclopedia of Portal Technologies and Applications

Arthur Tatnall
Victoria University, Australia

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Acquisitions Editor: Kristin Klinger
Development Editor: Kristin Roth
Senior Managing Editor: Jennifer Neidig
Managing Editor: Sara Reed
Assistant Managing Editor: Diane Huskinson
Copy Editors: Lanette Ehrhardt, April Schmidt, Katie Smalley, Angela Thor, and Larissa Vinci
Typesetter: Diane Huskinson and Laurie Ridge
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@idea-group.com
Web site: http://www.info-sci-ref.com

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanonline.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of portal technologies and applications / Arthur Tatnall, editor.
p. cm.
Summary: "This book offers complete coverage of the nature, characteristics, advantages, limitations, design, and evolution of Web portals. Other topics include semantic portals, philosophical portal issues, and personal portals. This authoritative encyclopedia encompasses the economics of setting up and using personal portals, knowledge management, strategic planning, user acceptance, security and the law"--Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-59140-989-2 (hardcover) -- ISBN 978-1-59140-990-8 (ebook)
1. Web portals--Encyclopedias. 2. World Wide Web--Encyclopedias. 3. Knowledge management--Encyclopedias. 4. Online information services--Encyclopedias. 5. Computer network resources--Encyclopedias. I. Tatnall, Arthur.
ZA4201.E53 2007
025.0403--dc22
2007007262

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Dynamic Taxonomies and Intelligent User-Centric Access to Complex Portal Information

Giovanni M. Sacco
Università di Torino, Italy

INTRODUCTION

One of the key requirements of portals is easy access to information, or *findability* according to Morville's definition (Morville, 2002). After a decade of using traditional access paradigms, such as queries on structured database systems and information retrieval or search engines, the feeling that "search does not work" and "information is too hard to find" is now reaching a consensus level. The problem is that traditional access paradigms are not suited to most search tasks, that are exploratory and imprecise in essence: the user needs to explore the information base, find relationships among concepts and think alternatives out in a guided way.

New access paradigms supporting exploration are needed. Since the goal is end-user interactive access, a holistic approach in which modeling, interface and interaction issues are considered together, must be used and will be discussed in the following.

BACKGROUND

Four retrieval techniques are commonly used: (a) information retrieval (IR) techniques (van Rijsbergen, 1979) recently dubbed search engines; (b) queries on structured databases; (c) hypertext/hypermedia links and (d) static taxonomies, such as Yahoo!.

The limitations of IR techniques are well known: a 1985 study reported that only 20% of relevant documents were actually retrieved (Blair & Maron, 1985). Such a significant loss of information is due to the extremely wide semantic gap between the user model (concepts) and the model used by commercial retrieval systems (words). Other problems include poor user interaction because the user has to formulate his query with no or very little assistance, and no exploration capabilities since results are presented as a flat list with no systematic organization. Database queries require structured data and are not easily applicable to situations, such as portals, in which most information is textual and not structured or loosely structured.

Hypermedia (see Groenbaek & Trigg, 1994) is quite flexible, but it gives no systematic picture of relationships among documents; exploration is performed one document at a time, which is quite time consuming; and building and maintaining complex hypermedia networks is very expensive.

Traditional taxonomies are based on a hierarchy of concepts that can be used to select areas of interest and restrict the portion of the infobase to be retrieved. Taxonomies support abstraction and are easily understood by end-users. However, they are not scalable for large information bases (Sacco, 2002), and the average number of documents retrieved becomes rapidly too large for manual inspection.

Solutions based on semantic networks, ontologies, and Semantic Web (Berners-Lee et al., 2001) are more powerful than plain taxonomies. However, general semantic schemata are intended for programmatic access, and are known to be difficult to understand and manipulate by the casual user. User interaction must be mediated by specialized agents, which increases costs, time to market, and decreases the transparence and flexibility of user access.

DYNAMIC TAXONOMIES

Dynamic taxonomies (Sacco, 1987, 1998, 2000, also called *faceted classification systems*) are a general knowledge management model based on a multidimensional classification of heterogeneous data items and are used to explore/browse complex information bases in a guided yet unconstrained way through a visual interface.

The intension of a dynamic taxonomy is a taxonomy designed by an expert. This taxonomy is a concept hierarchy going from the most general to the most specific concepts. Directed acyclic graph taxonomies modeling multiple inheritance are supported but rarely required. A dynamic taxonomy does not require any other relationships in addition to *subsumptions* (e.g., IS-A and PART-OF relationships).

In the extension, items can be freely classified under n (n>1) concepts at any level of abstraction (i.e., at any level in the conceptual tree). This multidimensional classification is a generalization of the mono-dimensional classification scheme used in conventional taxonomies and models common real-life situations. First, items are very often about different concepts: for example, a news item on September 11, 2001, can be classified under "terrorism," "airlines," "USA," and so forth. Second, items to be classified usu-

ally have different features, "perspectives" or facets (e.g., time, location, etc.), each of which can be described by an independent taxonomy.

In dynamic taxonomies, a concept C is just a label that identifies all the items classified under C. Because of the subsumption relationship between a concept and its descendants, the items classified under C (items(C)) are all those items in the *deep extension* of C, that is, the set of items identified by C includes the *shallow extension* of C (i.e., all the items directly classified under C) union the deep extension of C's sons. By construction, the shallow and the deep extension for a terminal concept are the same.

There are two important immediate consequences of this approach. First, since concepts identify sets of items, logical operations on concepts can be performed by the corresponding set operations on their extension. This means that the user is able to restrict the information base (and to create derived concepts) by combining concepts through the normal logical operations (and, or, not). Second, dynamic taxonomies can find all the concepts related to a given concept C: these concepts represent the conceptual summary of C. Concept relationships other than subsumptions are inferred through the extension only, according to the following *extensional inference rule*: two concepts, A and B, are related if there is at least one item, *d*, in the knowledge base which is classified at the same time under A or under one of A's descendants and under B or under one of B's descendants. For example, we can infer an unnamed relationship between *terrorism* and *New York*, if an item classified under *terrorism* and *New York* exists. At the same time, since *New York* is a descendant of *USA*, also a relationship between *terrorism* and *USA* can be inferred. The extensional inference rule can be seen as a device to infer relationships on the basis of empirical evidence.

The extensional inference rule can be easily extended to cover the relationship between a given concept C and a concept expressed by an arbitrary subset S of the universe: C is related to S if there is at least one item d in S, which is also in items(C). Hence, the extensional inference rule can produce conceptual summaries not only for base concepts, but also for any logical combination of concepts. Since it is immaterial how S is produced, dynamic taxonomies can produce summaries for sets of items produced by other retrieval methods such as database queries, shape retrieval, and so forth, and therefore access through dynamic taxonomies can be easily combined with any other retrieval method.

Dynamic taxonomies work on conceptual descriptions of items, so that heterogeneous items of any type and format can be managed in a single, coherent framework. Finally, since concept C is just a label that identifies the set of the items classified under C, concepts are language-invariant, and multilingual access can be easily supported by maintaining different language directories, holding language-specific labels for each concept in the taxonomy. If the metadata descriptors used to describe an item use concepts from the taxonomy, then also the actual description of an item can be translated on the fly to different languages.

Exploration

The user is initially presented with a tree representation of the initial taxonomy for the entire knowledge base. Each concept label has also a count of all the items classified under it, i.e., the cardinality of items(C) for all C's. The initial user focus F is the universe, i.e., all the items in the information base.

In the simplest case, the user selects a concept C in the taxonomy and *zoom* over it. The zoom operation changes the current state in two ways. First, concept C is used to refine the current *user focus* F, which becomes F∩items(C). Items not in the focus are discarded. Second, the tree representation of the taxonomy is modified in order to summarize the new focus. All and only the concepts related to F are retained and the count for each retained concept C' is updated to reflect the number of items in the focus F that are classified under C'. The *reduced taxonomy* is derived from the initial taxonomy by pruning all the concepts not related to F, and it is a conceptual summary of the set of documents identified by F, exactly in the same way as the original taxonomy was a conceptual summary of the universe. In fact, the term *dynamic taxonomy* indicates that the taxonomy can dynamically adapt to the subset of the universe on which the user is focusing, whereas traditional, static taxonomies can only describe the entire universe.

The retrieval process can be seen as an iterative thinning of the information base: the user selects a focus, which restricts the information base by discarding all the items not in the current focus. Only the concepts used to classify the items in the focus and their ancestors are retained. These concepts, which summarize the current focus, are those, and only those, concepts that can be used for further refinements. From the human computer interaction point of view, the user is effectively guided to reach his goal by a clear and consistent listing of all possible alternatives, and, in fact, this type of interaction is often called *guided thinning* or *guided navigation*.

Figures 1 to 5 show how the zoom operation works. Figure 1 shows a dynamic taxonomy: the upper half represents the intension with circles representing concepts; the lower half is the extension, and documents are represented by rectangles. Arcs going down represent subsumptions; arcs going up represent classifications. In order to compute all the concepts related to H, we first find, in Figure 2, all the documents classified under H (that is, the deep extension of H, items(H)) by following all the arcs incident to H (and, in general, its descendants): items(H) = { b, c, d }. All the items not in the deep extension of H (Figure 3) are removed from the extension. In Figure 4, the set of all the concepts

Figure 1. A dynamic taxonomy: the intension is above, the extension below. Arrows going down denote subsumptions, going up classification

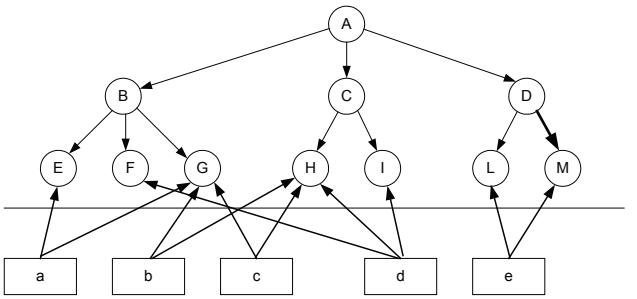


Figure 2. Focusing on concept H: finding all the items classified under H

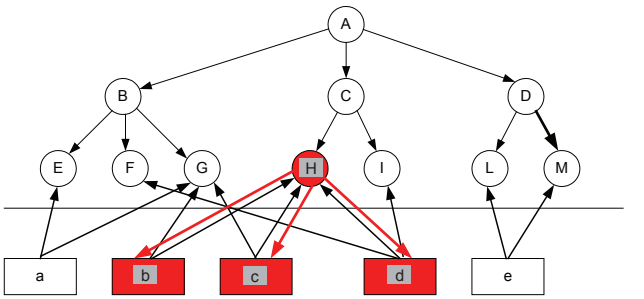


Figure 3. All the items not classified under H are removed

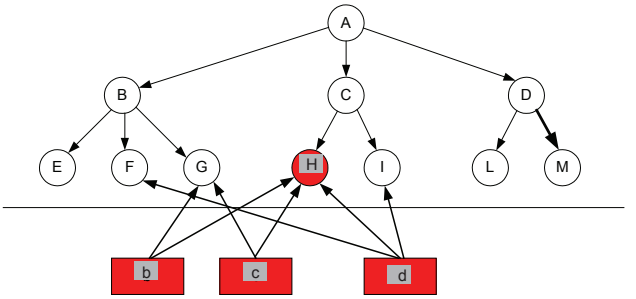


Figure 4. All the concepts under which the items in the focus are classified (and, because of subsumptions) their ancestors are related to H

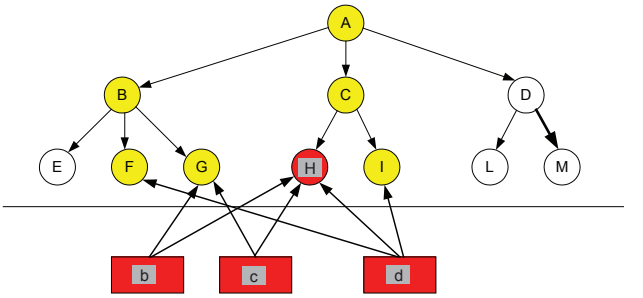
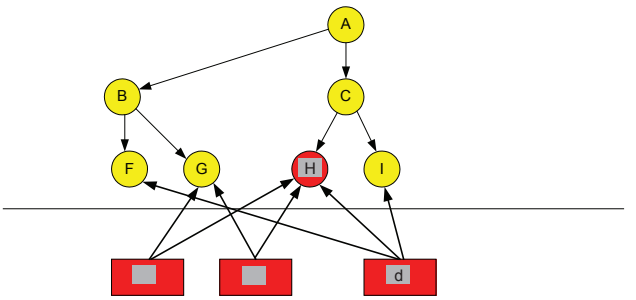


Figure 5. The reduced taxonomy: all concepts not related to the current focus are pruned



under which the documents in items(H) are classified, B(H), is found by following all the arcs leaving each element in the set: B(H) = { F, G, H, I }. The inclusion constraint implied by subsumption states that if items(C) denotes the set of documents classified under C and C' is a descendant of C in the taxonomy, items(C') $\subseteq$ items(C) (Sacco, 2000). This is equivalent to say that a document classified under C' is also classified under C. Hence, the set of concepts related to H is given by B(H) union all the ancestors of all the concepts in B(H), that is., the set of all concepts related to H is {F, G, H, I, B, C, A}. Finally, in Figure 5, all the concepts not related to H are removed from the intension, thus producing a reduced taxonomy that fully describes all and only the items in the current focus. A visual interaction example is provided in the article "E-commerce portals: guided product selection and comparison."

Advantages

The advantages of dynamic taxonomies over traditional methods are dramatic in terms of convergence of exploratory patterns and in terms of human factors. The analysis by Sacco (2002) shows that three zoom operations on terminal concepts are sufficient to reduce a 1,000,000 item information base

described by a compact taxonomy with 1,000 concepts to an average 10 items. Experimental data on a real newspaper corpus of over 110,000 articles, classified through a taxonomy of 1100 concepts, reports an average 1246 documents to be inspected by the user of a static taxonomy vs. an average 27 documents after a single zoom on a dynamic taxonomy.

Dynamic taxonomies require a very light theoretical background: namely, the concept of a taxonomic organization and the zoom operation, which seems to be very quickly understood by end-users. Hearst, English, Sinha, Swearingen, & Yee (2002) and Yee, Swearingen, Li, and Hearst (2003) conducted usability tests on a corpus of art images. Despite slow response times, access through a dynamic taxonomy was shown to produce a faster overall interaction and a significantly better recall than access through text retrieval. Perhaps more important are the intangibles: the feeling that one has actually considered all the alternatives in reaching a result. Although few usability studies exist, the widespread adoption by e-commerce portals, such as Yahoo!, Lycos, Bizrate, and so forth, empirically supports this initial evidence.

The derivation of concept relationships through the extensional inference rule has important implications on conceptual modeling. First, it simplifies taxonomy creation and maintenance. In traditional approaches, only the relationships among concepts explicitly described in the conceptual schema are available to the user for browsing and retrieval. Therefore, all possible relationships must be anticipated and described: a very difficult if not helpless task. In dynamic taxonomies, no relationships in addition to subsumptions are required, because concepts relationships are automatically derived from the actual classification. For this reason, dynamic taxonomies easily adapt to new relationships and are able to discover new, unexpected ones. Second, since dynamic taxonomies synthesize compound concepts, these need usually not be represented explicitly. This removes the main cause of the combinatorial growth of traditional taxonomies. Sacco (2000) developed guidelines that produce taxonomies that are compact and easily understood by users. Some are similar to basic faceted classification (Hearst et al., 2002; Ranganathan, 1965), at least in its basic form: the taxonomy is organized as a set of independent, "orthogonal" subtaxonomies (facets or perspectives) to be used to describe data. Although the term *faceted classification* is frequently used instead of *dynamic taxonomies*, it is a misnomer because (a) faceted classification only addresses conceptual modeling and very basic concept composition: conceptual summaries, reduced taxonomies and guided navigation are totally absent, and (b) faceted classification is a special case of the more general multidimensional classification on which dynamic taxonomies are built.

As an example of faceted design guidelines, consider a compound concept such as "19th century French paintings." It can be split into its facets: a location taxonomy (of which France is a descendant), a time taxonomy (of which the nineteenth century is a descendant) and finally an art taxonomy (of which painting is a descendant). The items to be classified under the compound concept will be classified under location>France, time>19th century and art>Painting instead. The extensional inference rule establishes a relationship among these concepts and the compound concept can be recovered by zooming on any permutation of them. In a conventional classification scheme, such as Dewey indexing (Dewey, 1996), in which every item is classified under a single concept, a number of different concepts equal to the Cartesian product of the terminals in the three taxonomies has to be defined. Such a combinatorial growth either results in extremely large conceptual taxonomies or in a gross conceptual granularity (Sacco, 2000). In addition, faceted design coupled with dynamic taxonomies makes it simple to focus on a concept, for example, 19th century, and immediately see all related concepts such as literature, painting, politics, and so forth, which are recovered through the extensional inference rule. In the compound concept approach, these correlations are unavailable because they are hidden inside the concept label.

Additional advantages include the uniform management of heterogeneous items of any type and format, easy multilingual access and easy integration with other retrieval methods. Dynamic taxonomies do not support reasoning beyond the extensional inference rule, and are therefore less powerful than general ontologies. However, they can be directly manipulated by users without the mediation of specialized agents and represent a quicker, less costly and more transparent alternative.

APPLICATIONS

The main industrial application is currently e-commerce. Assisted product selection is a critical step in most large-scale e-commerce systems (Sacco, 2003) and the advantages in interaction are so significant as to justify the restructuring of well-established e-commerce portals: current examples include Yahoo!, Lycos, Bizrate, and so forth.

However, dynamic taxonomies have an extremely wide application range and a growing body of literature indicates that their adoption benefits most portal applications. In addition to e-commerce, e-auctions and e-catalogs, key areas such as e-government portals (Sacco, 2005a, 2005c), human resources and job placement portals (Sacco, 2005c), news portals, art and museum portals (Hyvönen et al., 2004; Yee et al., 2003), medical guideline portals (Wollersheim & Rahayu, 2002) and diagnostic (Sacco, 2005b) and CRM (customer relationship management) portals, are being investigated and initial solutions deployed. An additional area is multimedia databases, where dynamic taxonomies can be used to integrate access by conceptual metadata and access

by primitive multimedia features (color, texture, etc.) into a single, coherent framework (Sacco, 2004). The reader is referred to the articles "E-commerce portals: guided product selection and comparison," and "Portals for integrated competence management" in this encyclopedia for an in-depth discussion of two important applications

A growing number of web-based commercial systems based on dynamic taxonomies exist. Among these, Knowledge Processors, Endeca, i411 and Siderean Software.

FUTURE TRENDS

Following the quick and widespread adoption by e-commerce portals, we expect dynamic taxonomies to become pervasive in the short period, and to replace or integrate traditional techniques in most portals. Current research is focused on three broad areas:

1. **Automatic Classification and Schema Design:** Dynamic taxonomies do not define how documents are actually classified. Current research focuses on automatic text classification (Dakka et al., 2005) and automatic classification from structured data (US Patent 6,763,349, 1998). Recent investigations (Sacco, 2005d) suggest that dynamic taxonomies can be automatically derived from semantically rich conceptual schemata and used as a user-centered front-end to complex information. Other research addresses the problem of specifying valid term compositions in faceted taxonomies for textual information (Tzitzikas et al., 2005)
2. **Extensions to the Model and Human Factors:** A fuzzy (Zadeh, 1965) classification, in which a document can be classified under several concepts with different probabilities, can sometimes be more appropriate than the boolean classification currently used (Sacco, 2004). Because of the holistic approach, human factors play a paramount role in devising extensions to the model and in critical issues such as the presentation and manipulation of the taxonomy, where several alternatives exist (see Yee et al., 2003 vs. Sacco, 2000, 2004).
3. **Centralized, Distributed, Federated Architectures:** The zoom operation and the subsequent reduction of the corpus taxonomy must be performed in real time because a slower execution would severely impair the sense of free exploration that the user of dynamic taxonomy systems experiences. Special data structures and evaluation strategies must be used (Sacco, 1998). In addition, distributed and federated architectures need to be investigated since centralized architectures are not always appropriate, because of organization needs and of performance and reliability bottlenecks.

CONCLUSION

Exploratory browsing applies to most practical situations and search tasks in portals: an extremely wide application range going from multilingual portals, to portals for e-commerce, e-auctions, e-government, human resources management, CRM, and so forth. In this context, dynamic taxonomies represent a dramatic improvement over other search and browsing methods, both in terms of convergence and in terms of full feedback on alternatives and complete guidance to reach the user goal. For these reasons, portals based on this paradigm are rapidly growing in number.

REFERENCES

Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May 17). The Semantic Web. *Scientific American,* 35-43.

Dakka, W., Ipeirotis, P. G., & Wood, K. R. (2005). Automatic construction of multifaceted browsing interfaces. In *Proceedings of the 14th ACM international Conference on information and Knowledge Management (CIKM '05*) (pp. 768-775).

Dewey, M. (1996). *Dewey decimal classification and relative index* (21st ed.) (J. S. Mitchell, J. Beall, W. E. Matthews, & G. R. New, Eds.) Dublin, OH: Forest Press.

Groenbaek, K., & Trigg R. (Eds.). (1994). Hypermedia. *Communications of the ACM, 37*(2).

Hearst, M., English, J., Sinha, R., Swearingen, K., & Yee, K.-P. (2002). Finding the flow in Web site search. *Communications of the ACM, 45*(9), 42-49.

Hyvönen, E., Saarela, S., & Viljanen, K. (2004). Application of ontology techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium (ESWS 2004)* (LNCS 3053, pp. 92-106). Springer-Verlag.

Morville, P. (2002). *The age of findability. Boxes and Arrows.* Retrieved April 29, 2002, from http://www.boxesandarrows.com/archives/002595.php.

Ranganathan, S. R. (1965). *The colon classification. Rutgers series on systems for the intellectual organization of information (Susan Artandi, Ed.)*. Piscataway, NJ: Rutgers University Press.

Sacco, G. M. (1987). Navigating the CD-ROM. In *Proceedings fo the International Conference On Business of CD-ROM.*

Sacco, G. M. (1998). Dynamic taxonomy process for browsing and retrieving information in large heterogeneous data bases, *US Patent 6,763,349*; also, Italian Patent 01303603.

Sacco, G. M. (2000). Dynamic taxonomies: A model for large information bases. *IEEE Transactions on Knowledge and Data Engineering 12*(2), 468-479.

Sacco, G. M. (2002). Analysis and validation of information access through mono, multidimensional and dynamic taxonomies. Tech Report Dept. of Informatica, University of Torino.

Sacco, G. M. (2003). The intelligent e-sales clerk: The basic ideas. In *Proc. INTERACT'03—9th IFIP TC13 International Conference on Human-Computer Interaction* (pp. 876-879).

Sacco, G. M. (2004). Uniform access to multimedia information bases through dynamic taxonomies, *IEEE 6th Intern. Symp. on Multimedia Software Engineering, (ISMSE'04*) (pp. 320-328).

Sacco, G. M. (2005a). No (e-)democracy without (e-)knowledge. In E-government: Towards electronic democracy. Int. Conf. IFIP TCGOV 2005, Bolzano (LNCS 3416, pp. 147-156).

Sacco, G. M. (2005b). Guided interactive diagnostic systems. *18th IEEE International Symposium on Computer-Based Medical Systems (CBMS '05)* (pp. 117-122).

Sacco, G. M. (2005c). Guided interactive information access for e-citizens. In EGOV05—Int. Conf. on E-Government, within the Dexa Conf. Framework (LNCS 3591, pp. 261-268).

Sacco, G. M. (2005d). Discount semantics: Modeling complex data with dynamic taxonomies, Tech. Rep, Università di Torino.

Tzitzikas Y., Analyti A., & Spyratos N. (2005). Compound term composition algebra: The semantics. *Journal on Data Semantics, 2*, 58-84.

van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.

Wollersheim, D., & Rahayu, W. (2002). Methodology for creating a sample subset of dynamic taxonomy to use in navigating medical text databases. *Proc. IDEAS 2002 Conf.* (pp. 276-284)

Yee, K.-P., Swearingen, K., Li, K., & Hearst, M. (2003). Faceted metadata for image search and browsing. *Proc. ACM CHI 2003* (pp. 401-408).

Zadeh, L. (1965). Fuzzy sets. *Information Control, 8*, 338-353.

D

KEY TERMS

Extension, Deep: Of a concept C, denotes the shallow extension of C union the deep extension of C's sons.

Extension, Shallow: Of a concept C, denotes the set of documents classified directly under C.

Extensional Inference Rule: Two concepts A and B are related if there is at least one item *d* in the knowledge base which is classified at the same time under A (or under one of A's descendants) and under B (or under one of B's descendants).

Facet: One of several top level (most general) concepts in a multidimensional taxonomy. In general, facets are independent and define a set of "orthogonal" conceptual coordinates.

Subsumption: A subsumes B if the set denoted by B is a subset of the set denoted by A ($B \subseteq A$)

Taxonomy: A hierarchical organization of concepts going from the most general (topmost) to the most specific concepts. A taxonomy supports abstraction and models subsumption (IS-A and/or PART-OF) relations between a concept and its father. Tree taxonomies can be extended to support multiple inheritance (i.e., a concept having several fathers).

Taxonomy, Monodimensional: Taxonomy where an item can be classified under a single concept only

Taxonomy, Multidimensional: Taxonomy where an item can be classified under several concepts

Taxonomy, Reduced: In a dynamic taxonomy, a taxonomy, describing the current user focus set F, which is derived from the original taxonomy by pruning from it all the concepts not related to F.

User Focus: The set of documents corresponding to a user-defined composition of concepts; initially, the entire knowledge base.

Zoom: A user interface operation, that defines a new user focus by OR'ing user-selected concepts and AND'ing them with the previous focus; a reduced taxonomy is then computed and shown to the user.