

## **DBWorld Xtended: Semantic Dissemination of Information through Dynamic Taxonomies**

**Giovanni Maria Sacco**

(Dipartimento di Informatica, Università di Torino  
sacco@di.unito.it)

**Abstract:** An integrated semantic dissemination system based on dynamic taxonomies is presented. The system supports conceptual information pull through an easily understood visual interface. A similar interface is used to express user interests at a conceptual level so that precise push strategies can be implemented. This system is currently used to manage the announcements coming from DBWorld, one of the best-known computer science research mailing lists, but it can be easily adapted to the dissemination needs of very diverse application areas ranging from e-government, to e-commerce, personalized news, etc.

**Keywords:** dynamic taxonomies, pull strategies, push strategies, semantic dissemination

**Categories:** H.3.3, H.5.2

### **1 Introduction**

We live in a world where the sheer quantity of available information and its rate of growth are rapidly becoming limiting factors as important as the lack of information was in the pre-internet era. In this context, intelligent information dissemination is bound to play an extremely important role in a large number of diverse application areas ranging from e-government, to e-commerce, personalized news, video-on-demand, research, etc.

The information overload has been managed so far by traditional methods such as text-retrieval, database queries and traditional taxonomies. We argue that most search tasks, and notably those typical of a knowledge worker, are exploratory and imprecise in essence: the user needs to explore the information base, find relationships among concepts and thin alternatives out in a guided way.

Examples of this type of access include the selection of the “right” product to buy, of a candidate for a job, but also finding the likely cause of a malfunction, past research relevant to one’s project etc. Indeed, exploratory access applies to an extremely wide range of practical situations. Traditional access methods are not helpful in this context, so that new access paradigms are required. Since the goal is end-user interactive access, a holistic approach, in which modelling, interface and interaction issues are considered together, must be used.

Dynamic taxonomies, described in the following, provide guided exploration of the knowledge base and form the backbone of the system described in this paper, which improves the quality and precision of a well-established announcement list, DBWorld, and supports both interactive pull access and proactive push dissemination. DBWorld is interesting in this context for several reasons. First, intelligent event

dissemination for the computer science community is important *per se*. Second, the availability of alternate representations of this same list, implemented with traditional technologies, allows to conduct large-scale usability studies and to gather user impressions and suggestions. This is especially important since the average user is a knowledgeable user anyway. In addition, the relative simplicity of the semantic content and of the size of the information base, makes qualitative measures of precision, recall and time to get results viable and meaningful. Finally, the interest of this solution transcends the application described here. In fact, the problems and solutions described here can be easily extended to a wide range of important dissemination applications, such as the ones described above.

## 2 The DBWorld mailing list

DBWorld is a mailing list [DBWorld] managed by Raghu Ramakrishnan, Department of Computer Sciences, University of Wisconsin. An ACM SIGMOD resource, DBWorld accepts and disseminates messages relevant to the data base research community, even though its scope seems to become larger with time. The bulk of these messages consists of conference announcements and calls for papers. DBWorld can be accessed in two basic ways. The first one is by “browsing”, i.e. reading a sequential list of announcements (Figure 1), ordered by message arrival. The second one is to subscribe to the list and receive by e-mail the announcements as they are inserted in the list. Subscription is by far the most used option, but it results in one’s mailbox being the unfocused target of several messages per day.

An alternate access path is provided by the Center for Information Processing at the University of Münster, Germany [DBWorld2]. They manage a list, derived from the original DBWorld list, that is searchable through a simple text-retrieval box, and can be ordered by deadline (Figure 2). After we completed our system, we learned of yet another access path to DBWorld, provided by Amund Tveit, Department of Computer and Information Sciences, Norwegian University of Science and Technology. This new list, called EventSeer [EventSeer] and in beta stage at the time of writing, presents the interface shown in figure 3. Its main improvement over the other offers seems to be in the classification of announcements, but access to the list content is made by a shallow alphabetic taxonomy. Only the original DbWorld list supports push strategies, albeit in a rather primitive and unfocused way.

Sent	Message Type	From	Subject	Deadline	Web Page
18-Jan-2005	conf. ann.	Davide Martinenghi	CFP LAAC 05: Logical Aspects and Applications of Integrity Constraints	4-Mar-2005	web page
18-Jan-2005	conf. ann.	Turan Philippe Ph.	2nd CFP: International Workshop on Database Interoperability	25-Feb-2005	web page
18-Jan-2005	conf. ann.	Xiaolong Yang	The Fourth International Conference on the Optical Internet (COIN'2005)	31-Jan-2005	web page
18-Jan-2005	conf. ann.	Mykola Pecherizky	Last Cfp: Special Track on Data Mining, IEEE CBMS 2005	7-Feb-2005	web page
18-Jan-2005	book ann.	Zongmin Ma	Call for Book Chapters: Soft		

Events					
talks		workshops		international conferences	
<input type="text"/> <input type="button" value="search"/> <input type="button" value="past conferences"/>					
This is a list of all future conferences:					
from	to	description	location	deadline	
Jan 22nd,2005	Jan 29th,2005	51st Annual Conference on Current Trends in Theory and Practice of Informatics (SOFSSEM'05) (homepage)	Liptovsky Jan, Slovakia	Aug 23rd,2004	
Jan 23rd,2005	Jan 29th,2005	Workshop on Graph Asymmetries 2005 (homepage)	Palmerston North, New Zealand	Jan 7th,2005	
Jan 24th,2005	Jan 29th,2005	FAH IEEE International Symposium and School on Advance Distributed	Guadalajara, Mexico	Nov 1st,2004	

Figure 1: DBWorld in browse mode: announcements are listed by arrival time

Figure 2: Searchable DBWorld: search plus ordering by deadline

Search in Call-For-Papers, examples of queries:  
bioinformatics 2005, web services 2006, "software engineering" and IBM Thomas Watson Research Center

Search CFPs

Navigation for Call-For-Papers on acronyms (e.g. DBPL2005 under d), countries (location of event and affiliations of program committee members (e.g. Iceland under f)), institutions (universities and research institutions (e.g. Columbia University under c)), pc members (or other persons/roles such as chair/invited speaker mentioned in a CFP (e.g. Professor Bebo White under b)) and (important) topics (e.g. "Topic Detection and Tracking" under t)

Acronyms	Countries	Institutions	PC Members	Topics
a, b, c, d, e,	a, b, c, d, e,	a, b, c, d, e,	a, b, c, d, e,	a, b, c, d, e,
f, g, h, i, j, k,	f, g, h, i, j, k,	f, g, h, i, j, k,	f, g, h, i, j, k,	f, g, h, i, j, k,

EventSeer

Beta

### Topics - starting with K

1. **Kdd Framework And Process** [Amazon]  
[potCFP139](#), [potCFP147](#), [potCFP3392](#), [potCFP3408](#),  
[potCFP3528](#), [potCFP3536](#), [potCFP4120](#), [potCFP4324](#).
2. **Kernel Methods** [Amazon]  
[potCFP3205](#), [potCFP3216](#), [potCFP3240](#), [potCFP3515](#),  
[potCFP3535](#), [potCFP3838](#), [potCFP3899](#), [potCFP4892](#),  
[potCFP4914](#), [potCFP877](#), [potCFP906](#).
3. **Key Management** [Amazon]  
[potCFP2621](#), [potCFP4477](#), [potCFP4893](#), [potCFP4907](#).
4. **Key Management And Key Recovery** [Amazon]  
[potCFP4333](#).
5. **Key Management For File Systems** [Amazon]

Figure 3: EventSeer: access by text-retrieval or primitive facets

Figure 4: EventSeer: Topics starting with K are expanded

### 3 Dynamic Taxonomies

Dynamic taxonomies [Sacco, 87, 98, 00] are a general knowledge management model for complex, heterogeneous information bases. It has been applied to very diverse areas, including multimedia databases [Sacco, 04], electronic commerce [Sacco, 03], and medical guidelines [Wollersheim, 02]. 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. 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 several topics at any level of abstraction (i.e. at any level in the conceptual tree). This multidimensional classification is a departure from the monodimensional classification scheme used in conventional taxonomies. Besides being a generalization of a monodimensional classification, a multidimensional classification models common real-life situations. First, an item is very rarely classified under a single topic. Second, items to be classified usually have different features, "perspectives" or facets (e.g. Time, Location, etc.), each of which can be described by an independent taxonomy.

By taking a "nominalistic" approach (concepts are defined by instances rather than by properties), 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, i.e. 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 consequences of our 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 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 IS-A are inferred through the extension only, according to the following *extensional inference rule*: two concepts A and B are related iff there is at least one item D in the infobase 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 a (unnamed) relationship between Michelangelo and Rome, if an item that is classified under Michelangelo and Rome exists in the infobase. At the same time, since Rome is a descendant of Italy, also a relationship between Michelangelo and Italy 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 iff 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. In addition, dynamic taxonomies can produce summaries for sets of items produced by other retrieval methods such as database queries, shape retrieval, etc. and therefore access through dynamic taxonomies can be easily combined with other retrieval methods.

Dynamic taxonomies can be used to browse and explore the infobase in several ways. The preferred implementation follows. The user is initially presented with a tree representation of the initial taxonomy for the entire infobase. 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 infobase).

In the simplest case, the user can then select 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 focus F, which becomes  $F = F \cap \text{items}(C)$ . 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 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* is used to indicate 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 then be seen as an iterative thinning of the information base: the user selects a focus, which restricts (thins out) 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.

Dynamic taxonomies are often improperly referred to as faceted classification access [Hearst, 02] [Yee, 02]. It is important to note that a) a traditional faceted classification [Ranganathan, 65] is not required by dynamic taxonomies, which only require that documents be classified under at least two concepts, and that b) the

concepts of systematic summaries and guided searches are completely absent from traditional faceted classification theory. A taxonomic design based on facets, i.e. independent coordinates in the conceptual space, was proposed in [Sacco, 00] as a design guideline, but it is by no means a necessary condition (as claimed in [Yee, 02]) or a sufficient one. In fact, it was shown [Sacco, 02] that naive access based on facets does not provide sufficient convergence in search.

Solutions based on semantic networks (e.g. [Schmeltz Pedersen, 93]) are being reconsidered in the current effort on ontologies and Semantic Web. Although more powerful and expressive than taxonomies, general semantic schemata are difficult to understand and manipulate by the casual user. They are better suited to programmatic access and user interaction must be mediated by specialized agents. This increases costs, time to market and decreases generality and flexibility of user access.

#### 4 The Dissemination Framework

As in the original DBWorld list, DBWorld Xtended [DBWorldX] supports both browsing and push strategies. Both access modes are based on dynamic taxonomies, so that users can work at a more comfortable conceptual level. The conceptual schema used to describe announcements follows the guidelines described in [Sacco, 00]. In short, we organize our conceptual space according “orthogonal”, independent coordinates: since dynamic concept composition is fully supported by dynamic taxonomies, this organization minimizes the number of required concepts, and makes the extensional inference rule more efficient.

In the present application, we chose a very simple structure and considered only the following facets:

- *Type of announcement*: conference, book, journal. For the moment being, position openings are not present.
- *Topic*: primary and secondary topics as indicated in calls for papers. Although there is a well-established index of computing literature, the ACM Computing Classification System [ACM-CSS], we decided for a new topic structure, which is less detailed and easier to navigate and has a better coverage for emerging interest areas in computer science.
- *Location*: this facet represents the geographical location for conferences; nations are grouped by continents.
- *Date*: this facet includes start and end dates for conferences, as well as the first deadline, which applies to all types of announcements.

A pull interaction is shown in figures 5 to 7. The user selects the first focus freely, according to his interests or needs. In figure 5, he selects Europe, and zooms. The reduced taxonomy shown in Figure 6 shows the conceptual summary of the focus, i.e. all the concepts related to Europe, by using the same taxonomic structure. Concepts not related to Europe are pruned from the taxonomy through the extensional inference rule. For instance, books and journals are pruned because they do not have a geographical location; topics that only occur for USA or Japan conferences are pruned as well. In Figure 7, after a subsequent zoom on Databases, the user expands the concept Software Engineering and reads the relevant documents.

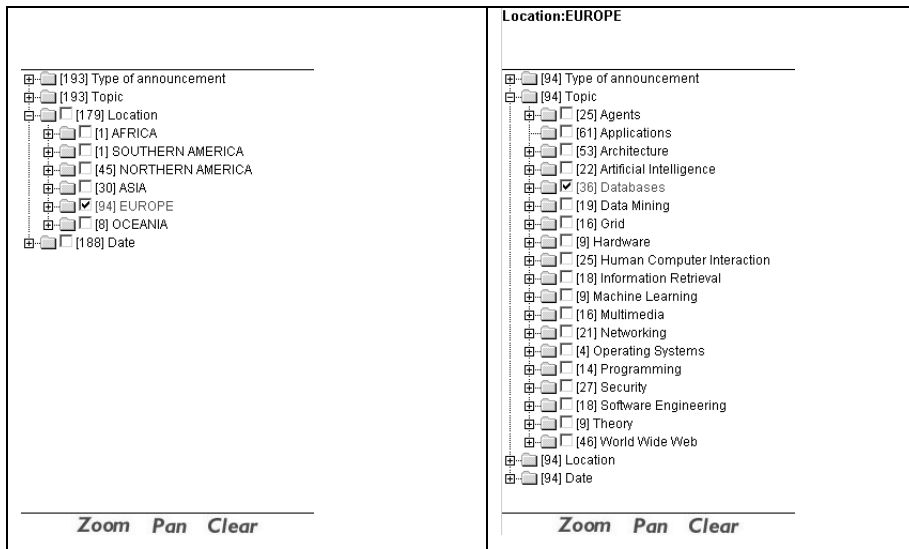


Figure 5: DbWorld Xtended: preparing to zoom on Europe

Figure 6: DbWorld Xtended: conceptual summary for Europe and preparing to zoom on Databases

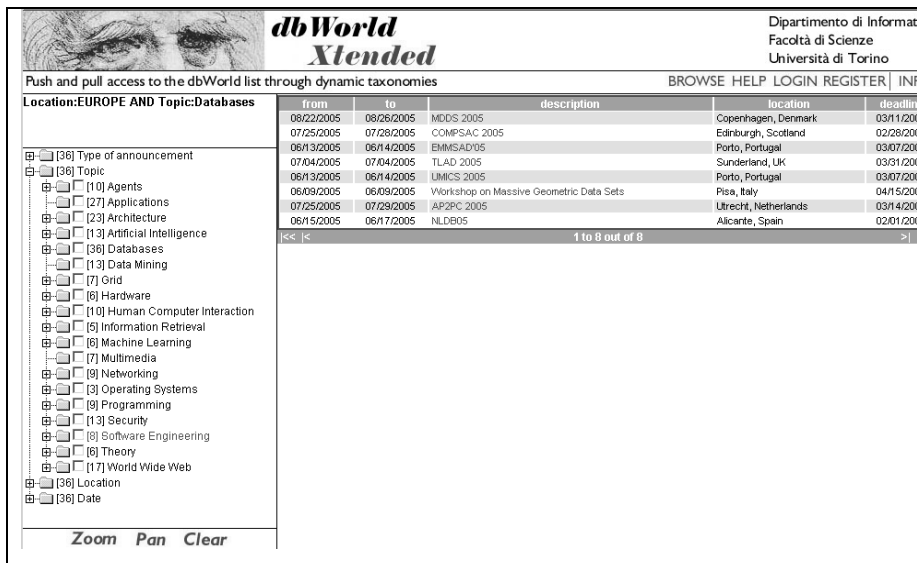


Figure 7: DbWorld Xtended: conceptual summary for Europe AND Database and exploring Software Engineering

The user was guided to reach his goal: at each step the systems provided a clear and understandable summary of all the possible focus alternatives. Just three steps

were required reduce a 193 item information base to just 8 items, even though general topics were used.

The advantages of dynamic taxonomies over traditional methods are dramatic in terms of convergence of exploratory patterns (see [Sacco, 02]) and in terms of human factors. 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. Yee et al. [Yee, 02] conducted usability tests on a corpus of art images. Despite an inefficient implementation that caused slow response times, their tests show that access through a dynamic taxonomy produced 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.

Specifications for push strategies are obtained on the basis of the same taxonomic conceptual schema used for the pull strategy. A specification can be issued by the user directly during his exploration. In the example above, the user can request to be mailed whenever new conferences in Europe on Databases are announced. An option allows retaining on the display all the concepts that would be pruned out by zooms, so that the user can specify push specifications even though no document currently satisfies them.

## 5 The architecture

DBWorld Xtended runs on Microsoft Windows 2003 Server and uses Knowledge Processors' Universal Knowledge Processor [UKP], a commercial system, to manage the dynamic taxonomy. The Universal Knowledge Processor is engineered to be easily integrated in existing software infrastructures and features extremely fast operations even on very large information bases. MySQL manages the document repository, which contains the actual text of call for papers, and the push subsystem. A java front-end assists registered users to create and send announcements. These are sent in XML in order to simplify transmission and checking by the list administration. Information retrieval access combined with dynamic taxonomies, although supported, is not currently implemented.

## 6 Conclusions and future work

DbWorld Xtended will go public in early 2005. The first, immediate goal is to make it an efficient site to serve the Computer Science community at large. The current limitation of DbWorld to database topics is obviously due to an attempt to reduce unfocused mails: this does not hold for our current implementation that can really become the single site where all computer science related events can be found. In turn, we believe that single repositories of events and more importantly of research publications are very needed in an area that is becoming more and more fragmented. From this point of view, we plan to integrate the announcement engine with access to document repositories where documents will be accessed by dynamic taxonomies. In parallel with conducting a large-scale usability study, we have short-term plans for using our solution for information dissemination in the context of e-government.

## Acknowledgements

DbWorld Xtended was built by Marco Lerda as a part of his Master thesis at the University of Torino. Marco Lerda also designed the conceptual schema. This work was funded in part by the Italian Ministry of Education, University and Research.

## References

- [ACM-CSS] ACM Computing Classification System, <http://www.acm.org/class/>
- [DBWorld] DBWorld mailing list, <http://www.cs.wisc.edu/dbworld/>
- [DBWorld2] searchable DBWorld list,  
<http://dbms.uni-muenster.de/menu.php3?page='events/index.php3?type=conference'>
- [DBWorldX] DBWorld Xtended, <http://dbworldx.di.unito.it>
- [EventSeer] the EventSeer mailing list, <http://eventseer.idi.ntnu.no/>
- [Hearst, 02] Hearst, M. et al., Finding the Flow in Web Site Search, *Comm. of the ACM*, 45: 9, 2002
- [Ranganathan, 65] Ranganathan, S. R., *The Colon Classification*. Rutgers Univ. Press, 1965
- [Sacco, 87] Sacco, G. M., Navigating the CD-ROM, *Int. Conf. Business of CD-ROM*, 1987
- [Sacco, 98] Sacco, G. M., Procedimento a tassonomia dinamica per il reperimento di informazioni su grandi banche dati eterogenee, Italian Patent 01303603; also US Patent 6,763,349
- [Sacco, 00] Sacco, G. M., Dynamic Taxonomies: A Model for Large Information Bases. *IEEE Transactions on Knowledge and Data Engineering* 12, 2, p. 468-479, 2000
- [Sacco, 02] Sacco, G. M., Analysis and Validation of Information Access through Mono, Multidimensional and Dynamic Taxonomies, Tech Report Dept. of Informatica, Univ. of Torino, 2002
- [Sacco, 03] Sacco, G. M., The Intelligent E-Sales Clerk: the Basic Ideas, in Krueger, H. & Rautenberg, M. (Eds.), *Proc. IFIP INTERACT'03 Conf. on Human-Computer Interaction*, 2003
- [Sacco, 04] Sacco, G. M., Uniform access to multimedia information bases through dynamic taxonomies, *IEEE Sixth Intern. Symp. on Multimedia Software Engineering*, Miami, FL, 2004
- [Schmelz Pedersen, 93] Schmelz Pedersen, G., A Browser for Bibliographic Information Retrieval, Based on an Application of Lattice Theory, *Proc. 1993 ACM SIGIR Conf.*, 1993
- [UKP] Knowledge Processors, Universal Knowledge Processor,  
[www.knowledgeprocessors.com](http://www.knowledgeprocessors.com)
- [Wollersheim, 02] 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.*, Edmonton, Canada, 2002
- [Yee, 02] Yee, K-P., et al., Faceted Metadata for Image Search and Browsing, *Proc. ACM CHI 2002*