INSTITUTE OF COMPUTER SCIENCE
FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS

INFORMATION SYSTEMS AND SOFTWARE TECHNOLOGY GROUP

The Semantic Index System

by

Panos Constantopoulos and Martin Doerr

The Semantic Index System (SIS) is a tool for describing and documenting large evolving varieties of highly interrelated data, concepts and complex relationships, as opposed to large homogeneous populations in fixed formats (handled by traditional DBMS). As such, it is suited for the representation of scientific knowledge and engineering designs or constructs. These kinds of data are also characterized by relative stability, i.e. they undergo few updates by comparison to, say, administrative, financial or observational data, which give rise to continuously changing sets of uniform items.

General Characteristics

The SIS consists of a persistent storage mechanism based on an object-oriented semantic network data model, and a generic interactive user interface to insert and retrieve information in various ways. Several automatic interfaces to other systems and customized user interfaces are provided.

The SIS offers significantly richer referencing mechanisms than relational or ordinary object oriented systems. Together with the very high query speed along references, these mechanisms allow to keep data and schema redundance-free; each notion has one identifier in the system, be at the schema, data, or field level. Fast built-in inheritance mechanisms gather, at query time, the relevant information per node from its environment.

By virtue of these properties, the system maintains an increasing amount of structural knowledge, needs minimal data input and does not cause consistency problems. On the other hand, it does not support logical formula evaluation, as expert systems do, to avoid the performance drawbacks associated, as well as problems in defining such formulas.

Structural knowledge, as in the SIS, is easily understood by non-experts. Moreover, it allows a straight-forward, highly efficient implementation, suitable for very large amounts of data. The SIS outperforms all relational systems on the market in lookup and traversal access times by a factor of about 25. Thus, it currently is a unique pragmatic solution for efficiently handling very large sets of highly structured data.

Main features

The main features of the SIS, listed below, result in a powerful combination of presentational flexibility, access efficiency and ease of adaptation.

- Uniform treatment of data and schema
  Data and schema are treated uniformly by the data entry, query system and user interface. This allows for dynamic schema definition and modification at runtime. Moreover, the schema is visible to the user, supporting the explanation and exploration of its structure.
- Powerful knowledge representation
  The data model offers the following general static knowledge representation mechanisms: unbounded instantiation hierarchy, multiple, strict inheritance (generalization), multiple classification and multivalued attributes which can have their own attributes.
- Domain specificity
  Items and properties are identified by logical names which, by their natural meaning, refer to the corresponding real world entities. Thus the data are presented in the specific terminology of the application domain and are easy to understand by experts of the domain without knowledge of the data model. By virtue of the ability to define the schema at runtime, the introduction of domain-specific concepts and terms is a simple interactive task.
- Effective, customizable user interface
  The user interface supports menu-guided and forms-based query formulation with graphical and textual presentation of the answer sets. It also supports graphical browsing in a hypertext-like manner. A hypertext annotation mechanism is also provided. Menu titles, menu layout and domain-specific queries are user-configurable. Thus the user interface can be customized to the application without changing the executable code.
  A forms-based interactive data entry facility is provided. It allows for entering data and schema information in a uniform manner. By employing the schema information, it automatically adapts itself to the structure of the various classes and subclasses. Furthermore, it is customizable to application-specific tasks, such as classification of items, addition of descriptive elements, etc.
- Network-oriented search mechanism
  The query system supports search by multiple and recursive conditions, as well as navigational search through the entire network of semantic relations.
- Multimedia data
  Any item in the SIS may reference a multimedia object, comprising images, video, sound or text, which are stored externally. The SIS recognizes such references and automatically generates calls to the appropriate presentation tools with the respective parameters, which result in a synchronous display of the multimedia object.

Standard features

- Transaction concept and concurrent multi-user access.
- Programmatic query interface
  The programmatic query interface offers a library of primitive access operations that enable external tools to transparently access the information of the SIS through their own user interfaces.
- Batch data import
  Data from external sources of information can be loaded into the SIS in ASCII (Telos) format. The same format can be generated by the SIS from interactively entered or loaded data (export).
- Report writing
  Query answer sets in textual or graphical (postscript) format can be saved in the standard file system for report writing purposes.

Performance

The current version of the SIS has been tested with a population of up to 850,000 entities and references. The test data contained 5 to 6 references per entity, hence about 150,000 objects in the terminology of object-oriented databases. The maximum capacity is 1 billion entities and references.

A recursive query on a binary tree including cycle detection with 1024 links requires about 2 seconds on a SUN SPARCstation. Up to a total population of 500,000, no significant influence of the population size on the query speed could be measured. Batch data import of 10,000 entities and references requires about 2 minutes on the same machine.

Integration with database systems and other tools

The SIS is designed to cooperate with external database systems and special-purpose tools, thus creating an integrated working environment. This integration is achieved by the following means :

- HP Toolbus command interface
  The SIS can receive HP Toolbus events to execute queries and to present the answer sets in any of the foreseen textual or graphical modes. Events can be sent from the SIS to other tools using particular query results as parameters. The latter is replaced by shell commands, if an HP environment is not available.
- The programmatic interface and batch data import, as presented above.
- Database cooperation
  A major application domain of the SIS is its usage as additional index to the more static part of data residing on an RDBMS, such as a systematic catalogue of the books in a library, whereas the more volatile data, such as circulation or accounting of a library, are handled by the RDBMS. There exists a SQL-based interface, which imports the minimal common description elements and the RDBMS identifiers into the SIS. All further descriptions, classifications and references are entered directly into the SIS. On retrieval, RDBMS identifiers are recognized and can be sent to the RDBMS for display or report purposes on the selected data.

Current applications

The SIS has already been incorporated in a software static analysis and class management tool, as well as in a museum information system.

The SIB Static Analyzer

The SIB Static Analyzer is a tool for the documentation and analysis of the static properties of software source code written in various programming languages to support the development process. So far, it has been employed with code written in COBOL, C++ and CooL.

As a unique feature, the SIB Static Analyzer allows analysis of programs using more than one programming language. All recursive relations are presented graphically in both directions, for instance, call trees and caller trees. The set of standard queries and presentations can be extended by the user to check possible side effects of modifications or to control programming rules. Other tools of the software production environment, as editors and debuggers, can be directly called from the Software Static Analyzer with suitable parameters.

Besides others, the system is currently used by Siemens-Nixdorf for analysing one of the greatest object oriented applications, with more than 2.5 million lines of code. Since all efforts to implement such static analyzers on standard databases have failed in the past due to performance reasons, it is the first such system providing persistency with a general purpose storage system and the reported capacity.

The SIB Class Management System

The SIB Class Management System provides access to object-oriented software libraries (class libraries). It allows the user to retrieve software objects by functional specification.

As extension of the Static Analyzer, the system is preloaded with the structural data of the respective objects. To make an object available, classification terms related to the functionality of the software objects are added interactively according to a faceted classification methodology, which is an extension of that proposed in the recent ESPRIT REBOOT project.

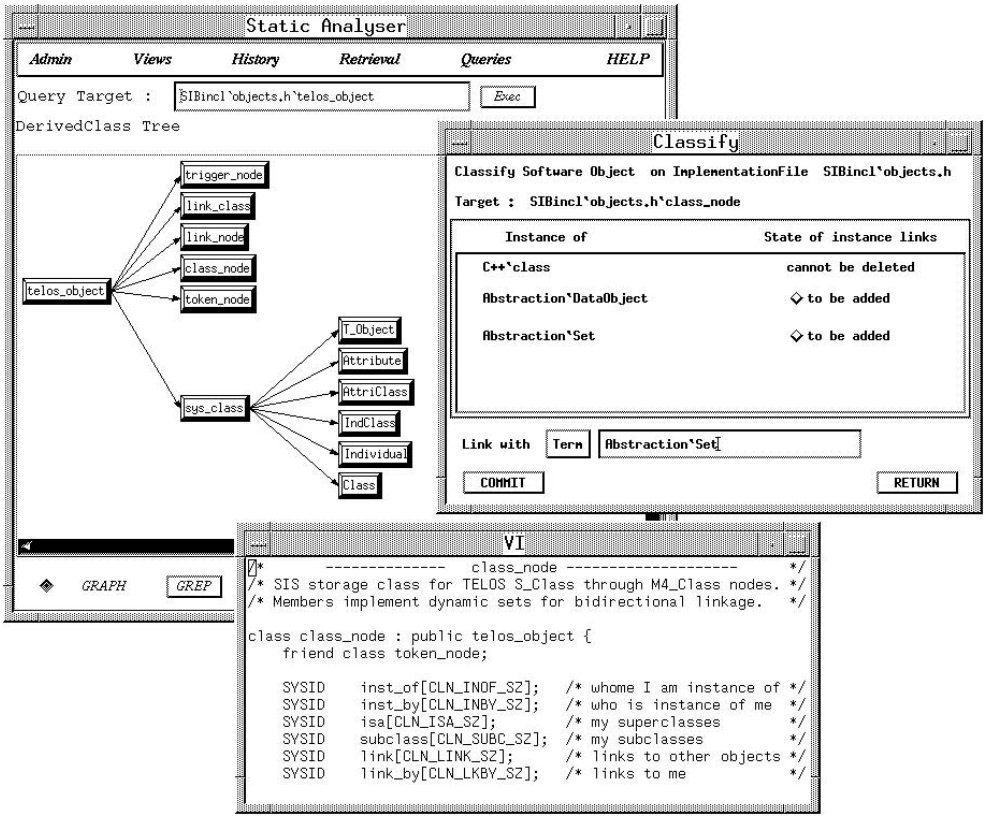
By employing a variety of inheritance rules, the data entry for classification is minimized and a maximum of consistency is maintained. The retrieval supports hierarchical as well as traditional keyword access. The views of structural and functional properties of software objects are fully integrated.

In comparison with other systems, the SIB Class Management System requires significantly less maintenance effort and provides excellent user support for retrieval specifications. The system was evaluated by Siemens Nixdorf, and is subject to license negotiations.

The CLIO Museum Information System

The CLIO Museum Information System is intended to serve as a scientific catalogue of museum artifacts, organized by collection, as opposed to the basic documentation and administrative purposes served by usual collections management systems.

By employing a special object-oriented cultural data model and the SIS, CLIO can store all the current knowledge, descriptive, contextual, historical, and other, about museum artifacts. CLIO can be used for documentation, research, and for the preparation of exhibitions, printed catalogues and educational material. It supports the causal linkage of



SIB Static Analyser and Class Management System

information; the storage of multiple, possibly contradictory values associated with their source and explanation; the use of a multitude of access points, such as temporal, geographical, cultural and historical context, physical properties, style, usage, etc; access by paths of successive references; and the storage and presentation of multimedia data.

The system interfaces to a conventional RDBMS-based collections management system. CLIO is being installed in museums in Greece. It is proposed for nation-wide usage in greek museums.

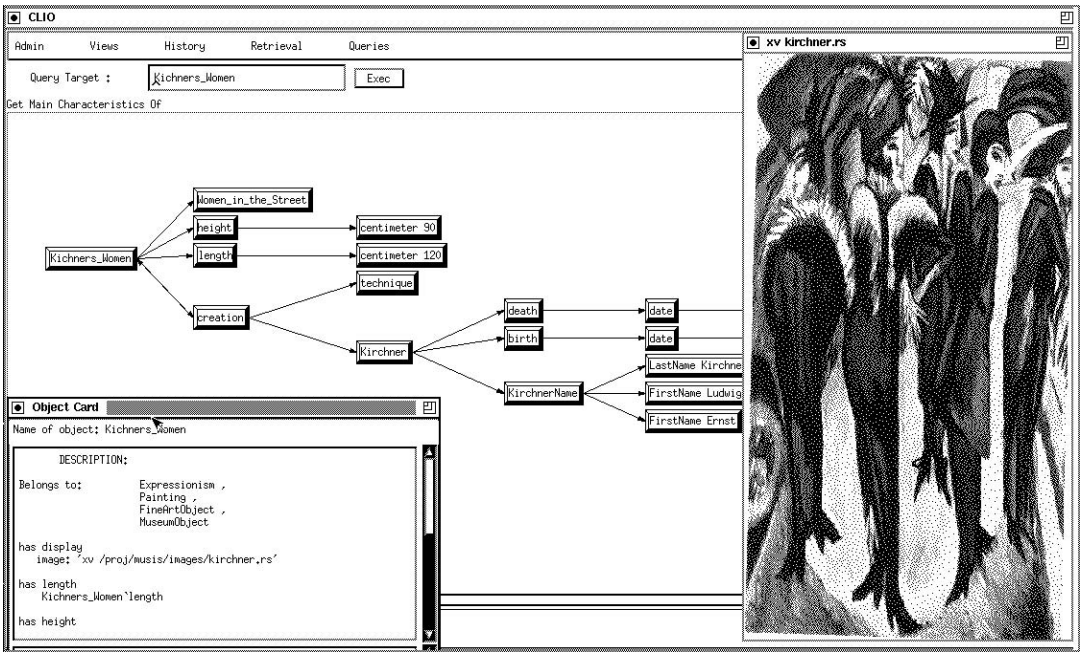
Competitive advantages

The advantages of the SIS concern usage, maintenance and performance.

The graphical user interface with various customizable and extensible views visualizes highly complex structures. The accessibility of schema information provides valuable explanations on the data presented. Moreover, the schema, at all levels of abstraction, is modifiable at runtime.

Redundance-free representation minimizes data input and consistency checking. Queries are supported on abstract or implicit properties, using domain specific terminology, which remain unaffected by modifications of the schema.

The system is capable of handling a very large number of items practically without decrease of performance.



CLIO : Documenting the cultural properties of a painting

The system can be integrated into an existing environment.

Installation environment

The SIS currently runs under UNIX System V, SUNOS, HP-UX 8.05, IBM AIX, SINIX D 5.41, and requires X11 Rel.4,5 and OSF Motif 1.1,1.2 .

A version running under Windows, Windows NT is under development.

Further information

Foundation of Research and Technology - Hellas
Institute of Computer Science
Information Systems and Software Technology Group
Science & Technology Park,
Vassilika Vouton, P.O. Box 1385, 71110
Heraklion, Crete, Greece.

Tel.: +30 (81) 391600, 391625 Fax: +30 (81) 391601, 391609
E-mail: infosys@ics.forth.gr