Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 60 (2007) 596–624
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Towards an intelligent database system founded on the SP theory of computing and cognition

J. Gerard Wolff *

CognitionResearch.org.uk, 18 Penlon, Menai Bridge, Anglesey LL59 5LR, UK

Received 15 March 2006; accepted 22 April 2006
Available online 16 May 2006

Abstract

The SP theory of computing and cognition, described in previous publications, is an attractive model for intelligent databases because it provides a simple but versatile format for different kinds of knowledge, it has capabilities in artificial intelligence, it can function effectively in the face of errors in its input data, and it can function like established database models when that is required.

This paper first describes the SP theory in outline and the computer models in which it is expressed. The main sections of the paper describe, with examples from the SP62 computer model, how the SP framework can emulate other abstract models used in database applications: the relational model (including retrieval of information in the manner of query-by-example, creating a join between two or more tables, and aggregation), object-oriented models (including class-inclusion hierarchies, part-whole hierarchies and their integration, inheritance of attributes, cross-classification and multiple inheritance), and hierarchical and network models (including discrimination networks). Comparisons are made between the SP model and those other models.

The artificial intelligence capabilities of the SP model are briefly reviewed: representation and integration of diverse kinds of knowledge in one versatile format; fuzzy pattern recognition and recognition at multiple levels of abstraction; best-match and semantic forms of information retrieval; various kinds of exact reasoning and probabilistic reasoning; analysis and production of natural language; planning; problem solving; and unsupervised learning. Also considered are ways in which current prototypes may be translated into an 'industrial strength' working system.

Keywords: Intelligent database; Information compression; Multiple alignment; Database model; Relational database; Object-oriented database; Hierarchical database; Network database

1. Introduction

The SP theory is a new theory of computing and cognition developed with the aim of integrating and simplifying a range of concepts in computing and cognitive science, with a particular emphasis on concepts in

* Tel.: +44 1248 712962.
E-mail address: jgw@cognitionresearch.org.uk
URL: www.cognitionresearch.org.uk

0169-023X/$ - see front matter
doi:10.1016/j.datak.2006.04.003

artificial intelligence. It is founded on the conjecture that all kinds of information processing, both natural (in brains) and artificial (in computers), may be understood as compression of information by the matching and unification of patterns. The name 'SP' expresses the idea that information compression promotes *Simplicity* in a body of information (by the removal of redundancy) whilst retaining as much as possible of its non-redundant descriptive *Power*.

The same two concepts lie at the heart of good science—where theories should, as far as possible, combine simplicity with descriptive or explanatory power—and they are a touchstone of success in engineering where, normally, we aim to create systems that are as simple as possible consistent with what they are designed to do. In the world of databases, these ideas were applied early on when it became apparent that, instead of hard coding each new database from scratch (each with its own user interface and methods for storing and retrieving information), an overall simplification of database systems could be achieved without loss of functionality or 'power' by developing a general-purpose 'database management system' that incorporates the features that are needed in every database (like those just mentioned) and could be loaded with different kinds of data according to need.

In a similar way, the SP project has sought to identify the elements that are shared by different aspects of intelligence (pattern recognition, information retrieval, reasoning, learning and so on) and to embody these shared elements in one general-purpose abstract model or 'system' for the representation and processing of knowledge. An overview of the SP theory is presented in [35]. More detail may be found in [37] and in earlier publications cited in those two sources.

1.1. The SP theory and intelligent databases

Amongst other things, the SP system provides an attractive model for database applications, especially those requiring a measure of human-like 'intelligence'. There is, of course, a wide variety of existing database systems that exhibit varying degrees and kinds of intelligence [10] and it is reasonable to ask what may be gained by creating yet another system in that domain. In brief, the main attractions of the SP model in this connection are that:

- It provides an extraordinarily simple yet versatile format for representing knowledge that facilitates the seamless integration of diverse kinds of knowledge.
- It provides a framework for processing that knowledge that integrates and simplifies a range of artificial intelligence functions including fuzzy pattern recognition and recognition at multiple levels of abstraction, best-match and semantic forms of information retrieval, various kinds of exact reasoning and probabilistic reasoning, analysis and production of natural language, planning, problem solving, and unsupervised learning.
- It can function effectively despite errors of omission, commission and substitution in incoming data.
- And it has the ability to emulate established database models when that is required.

As an aid to thinking and as a means of testing ideas, the abstract form of the SP system has been translated into the relatively concrete form of software simulations running on an ordinary computer, described in outline below (Section 2.3). A programme of further development will be needed to translate these prototypes into an 'industrial strength' working system (Section 7).

1.2. Databases and redundancies in knowledge

The SP system aims to compress any given body of knowledge by removing redundancies as far as possible. This is consonant with the generally-acknowledged principle that a 'well-structured' database should minimise redundancies. It is, for example, a recipe for confusion and inconsistencies if there are two different records in a database for 'John Smith' and the details of his address, telephone number and so on. However, it is clear that there is a need for redundancies in knowledge in the form of backup copies (as a safeguard against the risk of catastrophic loss of data) or mirror copies (to speed up processing when data is distributed across a network).

The resolution of this apparent contradiction is to say that, within each copy of a database, redundancies should, as far as possible, be minimised but that multiple copies may be made for the reasons indicated. For