

PACKAGE ID - 001118SUN0000 SID1.0

KWIC TITLE - System for Information Discovery

AUTHORS - Adams, K.J.
Pacific Northwest Lab., Richland, WA (United States)

Bohn, S.J.
Pacific Northwest Lab., Richland, WA (United States)

Crow, V.
Pacific Northwest Lab., Richland, WA (United States)

Harris, S.
Pacific Northwest Lab., Richland, WA (United States)

Koontz, A.
Pacific Northwest Lab., Richland, WA (United States)

Miller, N
Pacific Northwest Lab., Richland, WA (United States)

Nakamura, G.
Pacific Northwest Lab., Richland, WA (United States)

Pennock, K.
Pacific Northwest Lab., Richland, WA (United States)

Pottier, M.C.
Pacific Northwest Lab., Richland, WA (United States)

Thomas, J.
Pacific Northwest Lab., Richland, WA (United States)

Younkin, C.
Pacific Northwest Lab., Richland, WA (United States)

LIMITATION CODE -COPY **AUDIENCE CODE** - LIM

COMPLETION DATE - 03/26/1996 **PUBLICATION DATE** - 03/29/1996

DESCRIPTION - SID characterizes natural language based documents so that they may be related and retrieved based on content similarity. This technology processes textual documents, autonomously identifies the major topics of the document set, and constructs an interpretable, high dimensional representation of each document. SID also provides the ability to interactively re-weight representations based on user needs, so that users may analyze the data set from multiple points of view. The particular advantages SID offers are speed, data compression, flexibility in representation, and incremental processing.

PACKAGE CONTENTS - Media Directory; Software Abstract; Media Includes Source;

PACKAGE ID - 001118SUN0000 SID1.0

PACKAGE CONTENTS - (CONT)

SOURCE CODE INCLUDED? - Yes

MEDIA QUANTITY - 1 CD Rom

METHOD OF SOLUTION - SID is capable of processing data sets of arbitrary size. SID initially compresses the vocabulary associated with a data set, algorithmically selecting the most important words as features for further analysis. SID then algorithmically identifies the most distinguishable topics from the filtered set of words, and uses these topics as dimensions in a high dimensional vector representation of the information space. The resulting document representation can be used for querying and visualization.

COMPUTER - SUN

OPERATING SYSTEMS - SUN OS and SUN Solarus

PROGRAMMING LANGUAGES - C++

SOFTWARE LIMITATIONS - The system has only been tested on data sets as large as 40MB, though no a priori limit on the size of the data set exists.

SOURCE CODE AVAILABLE (Y/N) - Y

UNIQUE FEATURES - Text compression, interpretable vector representations, interactive characterization of documents based on user information needs.

HARDWARE REQS - SUN Sparc2 workstation or higher having a minimum 48MB of RAM, or equivalent Silicon Graphics workstation.

ABSTRACT STATUS - Submitted 10/16/96 Released AS-IS 12/3/96

SUBJECT CLASS CODE - P

KEYWORDS -

COMPUTER PROGRAM DOCUMENTATION
S CODES
INFORMATION RETRIEVAL
DATA PROCESSING
COMPUTER GRAPHICS
ALGORITHMS
INFORMATION THEORY

EDB SUBJECT CATEGORIES -

990200

SPONSOR - DOE/ER

PACKAGE TYPE - AS - IS