

**PACKAGE ID** - 001245SGIIP00 SID/SPIRE

**KWIC TITLE** - System for Information Discovery

**AUTHORS** - Crow, V.

Pacific Northwest National Lab., Hanford, WA (United States)

Nakamura, G.

Pacific Northwest National Lab., Hanford, WA (United States)

Younkin, C.

Pacific Northwest National Lab., Hanford, WA (United States)

Hetzler, B.

Pacific Northwest National Lab., Hanford, WA (United States)

McQuerry, D.

Pacific Northwest National Lab., Hanford, WA (United States)

**LIMITATION CODE** -COPY

**AUDIENCE CODE** - LIM

**COMPLETION DATE** - 01/26/1998

**PUBLICATION DATE** - 01/26/1998

**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 reweight representations based on user need, so users may analyze the dataset from multiple points of view. The particular advantages SID offers are speed, data compression, flexibility in representation, and incremental processing. SPIRE consists of software for visual analysis of text-based information sources. This technology enables users to make discoveries about the content of very large sets of textual documents without requiring the user to read or presort the documents. It employs algorithms for text and word proximity analysis to identify the key themes within the documents. The results of this analysis are projected onto a visual spatial proximity display (Galaxies or Themescape) where document proximity represents the degree of relatedness of theme.

**PACKAGE CONTENTS** - Media Directory; Software Abstract; Supplementary Information (2 pages); Media Includes Executable Module, User's Guide, Sample Problem Input Data, Installation Instructions;

**SOURCE CODE INCLUDED?** - Yes

**PACKAGE ID** - 001245SGIIP00 SID/SPIRE

**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 to 10% of the original size, algorithmically selecting the most important words as features for further analysis. SID then algorithmically identifies the most distinguishable topics from the filtered set of word, and uses these topics as dimensions in a high dimensional vector representation of the information space. All documents are then characterized based on their values for each of the dimensions. Documents that are similar in meaning are likely to be proximal in the vector space. The resulting document representation can be used for querying and visualization. SPIRE takes the vector representations of document context that have been created by SID or other text analysis engine, clusters the documents based on content similarity, and projects the clustered documents into a two-dimensional representation for easy interaction. SPIRE Version 3.0 includes three different user-selectable algorithms for clustering, all based on different statistical techniques.

**COMPUTER** - SILICON GRAPHIC

**OPERATING SYSTEMS** - Irix 5.3 or higher, SUN Solaris 2.5 or later

**PROGRAMMING LANGUAGES** - C++

**SOFTWARE LIMITATIONS** - The software has been tested on datasets as large as 650MB.

**SOURCE CODE AVAILABLE (Y/N)** - Y

**UNIQUE FEATURES** - Text compression, interpretable vector representation, interactive characterization of documents based on user information needs. Vastly improved performance over previous version of the software enables its practical application to much larger document sets, while the file format definition and subsetting capabilities add to SID's ability to successfully deal with a wider range of user problems.

**RELATED SOFTWARE** - Used as the standard text analysis engine for SPIRE (Spatial Paradigm for Information Retrieval and Exploration) Version 3.0 which is also included with this package.

**OTHER PROG/OPER SYS INFO** - The analysis results are only as good as the information relationships with the constituent documents in the dataset.

**HARDWARE REQS** - Minimum hardware requirements SUN Ultra 1 or better (SUN Ultra 2 with 3D Creator Graphics preferred) or SGI Indy or better. Requires 64MB of RAM; 128 MB preferred.

E S T S C  
ENERGY SCIENCE & TECHNOLOGY SOFTWARE CENTER  
SOFTWARE ABSTRACT

PAGE 3

DATE 03/12/2002

**PACKAGE ID** - 001245SGIIP00 SID/SPIRE

**HARDWARE REQS - (CONT)**

**ABSTRACT STATUS** - Released AS-IS 10/06/1998

**SUBJECT CLASS CODE** - Z

**KEYWORDS** -

COMPUTER PROGRAM DOCUMENTATION

S CODES

INFORMATION RETRIEVAL

**EDB SUBJECT CATEGORIES** -

990200

**SPONSOR** - DOE/DP

**PACKAGE TYPE** - AS - IS