

Using Superimposed and Context Information to Find and Re-find Sub-documents

Sudarshan Murthy

Department of CS, Portland State University

Uma Murthy, Edward Fox

Department of CS, Virginia Tech

Sub-document Search

- Users frequently search for sub-documents, but search engines tend to return list of documents
 - Not much indication of why the returned documents are relevant, and which parts of a document are relevant
- Some search engines include *snippets*, but provide no assistance to identify sub-documents within a document
 - Users routinely open documents and search *intra-document* for relevant sub-documents
- Finding effort is often repeated when re-finding

Windows Desktop Search

The screenshot shows a Windows Desktop Search window titled "environmental assessment - Windows Desktop Search". The search term "environmental assessment" is entered in the search bar and is circled in red. The window displays a list of search results with columns for Title, Author, and Date. The results include various documents related to environmental assessment, such as "mh2000_2b.pdf", "ICDE04-DP.doc", and "Geology EC 10_13_2000.doc". A red bracket on the right side of the window highlights the list of results, with the text "Documents returned" written next to it.

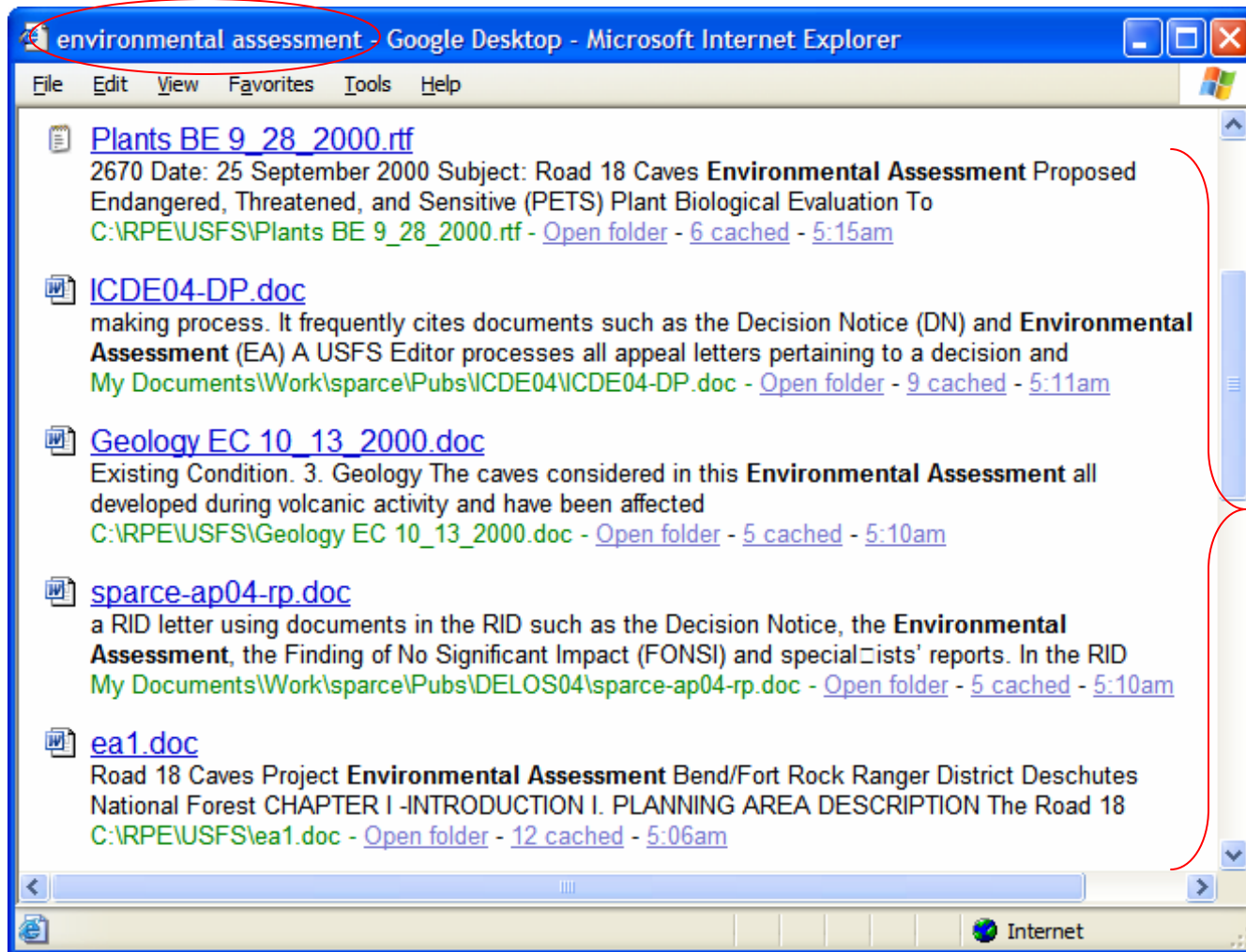
Search term

Documents returned

Title	Author	Date
mh2000_2b.pdf (ax03228n.aw)	USDA Forest Servi...	12/3/2001
ICDE04-DP.doc (Demo Proposal: Superimposed Applicat...)	Sudarshan Murthy	7/1/2003
ICDE04-DP.pdf (Microsoft Word - ICDE04-DP.doc)	smurthy	7/1/2003
archyCaves EA input#3.doc (SPEED MEMO)	BFR HERITAGE PR...	5/21/2002
Geology EC 10_13_2000.doc (RE:)	PCxx	5/21/2002
qresult.txt.html		4/5/2005
mh2000_2a.pdf (ax03228i.aw)	USDA Forest Servi...	12/3/2001
Spring 03.doc (Report to Student Program Committee (...))	SMurthy	3/4/2003
mh2000_2a.pdf (ax03228i.aw)	USDA Forest Servi...	12/3/2001
mh2000_2a.pdf (ax03228i.aw)	USDA Forest Servi...	12/3/2001
SPARCE.doc (SPARCE: Superimposed Pluggable Archite...)	Sudarshan S. Murt...	5/5/2003
Archy Report 9_18-2000.doc (SPEED MEMO)	BFR HERITAGE PR...	5/21/2002

Google Desktop

Search term



Documents returned

An Alternative: *Super Search*

- See list of documents and sub-documents along with detailed context information for sub-documents
 - Content: Sentence, paragraph, row, column, ...
 - Location: Page number, line number, sheet name, ...
- See sub-documents in context: Takes users *directly* to sub-documents; does not require authors to identify sub-document units
- Alternatively, browse context information without clicking through to sub-documents

Document Results*

Search term

The screenshot shows a window titled "Super Search (Via Google Desktop)". The search bar contains the text "environmental assessment" and a "Search" button. Below the search bar, there are five search results, each with a file name, a brief description, and a file path. The results are:

- [Plants BE 9 28 2000.rtf](#) [_Open in new window](#)
2670 Date: 25 September 2000 Subject: Road 18 Caves **Environmental Assessment** Proposed Endangered, Threatened, and Sensitive (PETS) Plant Biological Evaluation To: Wayne Gammon
C:\RPE\USFS\Plants BE 9_28_2000.rtf
- [ICDE04-DP.doc](#) [_Open in new window](#)
making process. It frequently cites documents such as the Decision Notice (DN) and **Environmental Assessment** (EA) A USFS Editor processes all appeal letters pertaining to a decision and
C:\Documents and Settings\smurthy\My Documents\Work\sparce\Pub...\ICDE04-DP.doc
- [Geology EC 10 13 2000.doc](#) [_Open in new window](#)
00 Chapter 1. Existing Condition. 3. Geology The caves considered in this **Environmental Assessment** all developed during volcanic activity and have been affected by subsequent
C:\RPE\USFS\Geology EC 10_13_2000.doc
- [sparce-ap04-rp.doc](#) [_Open in new window](#)
a RID letter using documents in the RID such as the Decision Notice, the **Environmental Assessment**, the Finding of No Significant Impact (FONSI) and specialists' reports. In the RID
C:\Documents and Settings\smurthy\My Documents\Work\sparc...\sparce-ap04-rp.doc
- [decision_notice2.doc](#) [_Open in new window](#)
Significant Impact For Road 18 Caves Project **Environmental Assessment** Deschutes National Forest Bend-Fort Rock Ranger District Deschutes County, Oregon Location The Road 18
C:\RPE\USFS\decision_notice2.doc

At the bottom of the window, it says "Ready." and "46 documents retrieved".

Documents returned

Sub-document Results*

The screenshot shows a window titled "Super Search (Via Google Desktop)". The search bar contains the text "environmental assessment" and a "Search" button. Below the search bar, there are two search results. The first result is for "Plants BE 9 28 2000.rtf" and the second is for "ICDE04-DP.doc". Both results show a snippet of text and a file path. The second result is expanded to show two numbered items, each with a "Page" and "Line" reference, a "See here:" link, and a snippet of text. The first item is "Page 3, Line 43" and the second is "Page 3, Line 83". Both items have links for "Sentence", "Paragraph", and "Line". The status bar at the bottom of the window indicates "Ready." and "2 sub-documents retrieved".

environmental assessment

Plants BE 9 28 2000.rtf [Open in new window](#)
2670 Date: 25 September 2000 Subject: Road 18 Caves **Environmental Assessment** Proposed Endangered, Threatened, and Sensitive (PETS) Plant Biological Evaluation To: Wayne Gammon
C:\RPE\USFS\Plants BE 9_28_2000.rtf

ICDE04-DP.doc [Open in new window](#)
making process. It frequently cites documents such as the Decision Notice (DN) and **Environmental Assessment** (EA) A USFS Editor processes all appeal letters pertaining to a decision and
C:\Documents and Settings\smurthy\My Documents\Work\sparce\Pub...\ICDE04-DP.doc

1. **Page 3, Line 43** See here: [Sentence](#) [Paragraph](#) [Line](#)
Sentence: It frequently cites documents such as the Decision Notice (DN) and **Environmental Assessment** (EA).
[See in context](#) [Browse context](#) [Remember](#)
2. **Page 3, Line 83** See here: [Sentence](#) [Paragraph](#) [Line](#)
Paragraph: Figure 3 shows a RIDPad instance (on the left) with information concerning the "Road 18 Caves" decision (taken in the Pacific Northwest Region of USFS). The instance shown has eight items in four groups. The group titled "**Environmental Assessment**" contains two groups. The information in the instance shown comes from three distinct base documents in two different base applications. (The item labeled "Comparison of Issues" contains an MS Excel mark; all other items contain MS Word marks.) All items were created using base-layer support included in the current implementation of SPARCE.
[See in context](#) [Browse context](#) [Remember](#)

Geology EC 10 13 2000.doc [Open in new window](#)
00 Chapter 1. Existing Condition. 3. Geology The caves considered in this **Environmental Assessment** all developed during volcanic activity and have been affected by subsequent
C:\RPE\USFS\Geology EC 10 13 2000.doc

Ready. 2 sub-documents retrieved

Search term

Some context information shown
- Location info
- Containing sentence and containing paragraph

2 sub-documents returned

See Detailed Context Information

The screenshot displays a web browser window titled "Super Search (Via Google Desktop)". The search query is "environmental assessment". The search results include:

- Plants BE 9_28_2000.rtf** (Open in new window)
2670 Date: 25 September 2000 Subject: Road 18 Caves **Environmental Assessment** Proposed Endangered, Threatened, and Sensitive (PETS) P...
C:\RPE\USFS\Plants BE 9_28_2000.r
- ICDE04-DP.doc** (Open in new window)
making process. It frequently cites o...
Assessment (EA) A USFS Editor pr...
C:\Documents and Settings\smurthy...
- Geology EC 10_13_2000.doc** (Open in new window)
00 Chapter 1. Existing Condition. 3. Geology The caves considered in this **Environmental Assessment** all developed during volcanic activity and have been affected by subsequent
C:\RPE\USFS\Geology EC 10_13_2000.doc

The "Context Browser (Mark Context) (Containment/Containing Pa..." window is overlaid on the search results. It features a menu bar (File, Edit, View) and a toolbar with icons for Save, Print, and Copy. The window is divided into two main sections:

- Context kinds and elements:** A tree view showing a hierarchy of context types: Content (Text, Formatted Text, HTML, Picture), Placement, Information, Presentation, Substructure, Containment (Containing Paragraph, Containing Section).
- Value of the context element currently selected:** A text area displaying the content of the selected context element. The text reads: "Figure 3 shows a RIDPad instance (on the left) with information concerning the 'Road 18 Caves' decision (taken in the Pacific Northwest Region of USFS). The instance shown has eight items in four groups. The group titled 'Environmental Assessment' contains two groups. The information in the instance shown comes from three distinct base documents in two different base applications. (The item labeled 'Comparison of Issues' contains an MS Excel mark; all other items contain MS

The "Browse context" link in the search results is circled in red.

Ready. 2 sub-documents retrieved

See a Sub-document in Context

environmental assessment Search Remember sub-documents seen in context

ICDE04-DP.doc Open in new window

making process. It frequently cites documents such as the Decision Notice (DN) and **Environmental Assessment (EA)** A USFS Editor processes all appeal letters pertaining to a decision and

C:\Documents and Settings\smurthy\My Documents\Work\sparce\Pub... \ICDE04-DP.doc

- Page 3, Line 43** See here: [Sentence](#) [Paragraph](#) [Line](#)
Sentence: It frequently cites documents such as the Decision Notice (DN) and **Environmental Assessment (EA)**.
[See in context](#) [Browse context](#) [Remember](#)
- Page 3, Line 83** See here: [Sentence](#) [Paragraph](#) [Line](#)
Paragraph: Figure 3 shows a RIDPad instance (on the left) with information concerning the "Road 18 Caves" decision (taken in the Pacific Northwest Region of USFS). The instance shown has eight items in four groups. The group titled "**Environmental Assessment**" contains two groups. The information in the instance shown comes from three distinct base documents in two different base applications. (The item labeled "Comparison of Issues" contains an MS Excel mark; all other items contain MS Word marks.) All items were created using base-layer support included in the current implementation of SPARCE.
[See in context](#) [Browse context](#) [Remember](#)

Go directly to a sub-document

Figure 3: A RIDPad instance and the Context Browser

Figure 3 shows a RIDPad instance (on the left) with information concerning the "Road 18 Caves" decision (taken in the Pacific Northwest Region of USFS). The instance shown has eight items in four groups. The group titled "**Environmental Assessment**" contains two groups. The information in the instance shown comes from three distinct base documents in two different base applications. (The item labeled "Comparison of Issues" contains an MS Excel mark; all other items contain MS Word marks.) All items were created using base-layer support included in the current implementation of SPARCE.

RIDPad affords many operations for items and groups. A user can create new items and groups, and move items between groups. The user can also rename, resize, and change visual characteristics such as color and font for

Ready. 2 sub-documents retrieved

Re-finding Sub-documents

Information "found" earlier is ranked higher

Context information can help re-find information

The screenshot shows a window titled "Super Search (Via Google Desktop)". The search bar contains "environmental assessment" and a "Search" button. A checkbox labeled "Remember sub-documents seen in context" is present. The results are divided into "Past Results" and "New Results".

Past Results

- [ea1.doc](#) [Open in new window](#)
Assessment Bend/Fort Rock Ranger District Deschutes National Forest CHAPTER I - INTRODUCTION I. PLANNING AREA DESCRIPTION The Road 18 Caves **Environmental Assessment**
C:\RPE\USFS\ea1.doc
- [sparce-ap04-rp.doc](#) [Open in new window](#)
thesizes a RID letter using documents in the RID such as the Decision Notice, the **Environmental Assessment**, the Finding of No Significant Impact (FONSI) and specialists' reports. In the RID
C:\Documents and Settings\smurthy\My Documents\Work\sparc...\sparce-ap04-rp.doc

1. Page 2, Line 46 See here: [Sentence](#) [Paragraph](#) [Line](#)
Line: **Environmental Assessment**, the Finding of No Significant
[See in context](#) [Browse context](#)

2. Page 2, Line 80 See here: [Sentence](#) [Paragraph](#) [Line](#)
Sentence: The instance shown has eight items (labeled Summary, Details, Comparison of Issues, Alternative A, Alternative B, Statement, Details, and FONSI) in four groups (labeled **Environmental Assessment**, Proposed Action, Other Alternatives, and Decision).
[See in context](#) [Browse context](#)

3. Page 2, Line 82 See here: [Sentence](#) [Paragraph](#) [Line](#)
Sentence: The group labeled "**Environmental Assessment**" contains two other groups.
[See in context](#) [Browse context](#) [Remember](#)

New Results

Ready. 3 sub-documents retrieved

Super Search Benefits

- Judge relevance more easily
 - See sub-documents, not just documents
- Reduce user click-through operations
 - Context information provides insight into sub-documents
- Reduce effort to find sub-documents in context
 - No need for user to perform *intra-document* search
- Leverage finding effort to re-find
 - See found results first; easy to remember found info.

Discussion Points

- Current search engine techniques can be inefficient when supporting finding and re-finding of sub-documents in this manner
 - Heterogeneous information
 - Showing information in context
 - How much of context information can a traditional inverted index provide?
 - Need to distinguish paragraphs, lines, rows, columns, page, ...