Dialog in a dot com labyrinth: Text based information retrieval in a graphical user interface culture.

James J. Heinis
The Dialog Corporation
11000 Regency Parkway #10
Cary, NC 27511
Chemical information

- Chemical information includes a wide range of topics:
  - Chemical Properties
  - E.g. Reactions, Synthesis
  - Physical Properties
  - Chemical Safety, Toxicological information.
  - E.g. MSDS
Record Structure

- Dialog and other Database aggregators:
- Field Oriented Databases: Document structure is uniform which enables efficient searches and standardized retrieval.
- Internet documents: Document structure varies between documents.
Search Strategies and Retrieval are based on Record Structure

- Commercial Text Based Retrieval Systems
  - e.g. Dialog uses a standardized structure.

- Web based information retrieval use HTML structure. Record structure is not as standardized.
  - HTML 3.2 (1996-1997)
  - HTML 4.01 (1999-now)
  - (There are also Netscape, and Internet Explorer browser specific parameters)

- This structure is adhered to by the information provider (creates database and is responsible for content)
Typical Record Description (Dialog)

- **FN**: DIALOG(R)File 2:INSPEC | … (Portions of Record Omitted)
- **AZ**: <DIALOG> 7127465 | <INSPEC> A2002-03-8160-006^B2002-01-0560-016 |
- **TI**: Photoluminescence in polypropylene induced by ultraviolet photons |
- **AU**: Ito, T.,^Toyoda, T.,^Hirai, N.,^Ohki, Y. | <CS1> Waseda Univ., Tokyo, Japan |
- **JN**: Transactions of the Institute of Electrical Engineers of Japan, Part A |
- **SO**: <VO> vol.121-A, no.9 | <PG2> 865-71 | <PU2> Inst. Electr. Eng. Japan |
- **PY**: Sept. 2001 | CP- Japan | CO- DGKRA8 | SN- 0385-4205 |
- **IC**: 0385-4205(200109)121A:9L.865:PPIU;1-K | MI- T195-2001-010 |
- **LA**: Japanese | DT- Journal Paper (JP) | TC- Experimental (X) |
- **AB**: We have measured photoluminescence (PL) spectra (Portion Omitted) |
- **RF**: 12 | **SF**: A^B | **NOTE FIELD NAME IS YELLOW.** |
- **DE**: impurities | organic insulating materials | oxidation | photochemistry | photoluminescence | polymer films | ultraviolet radiation effects |
- **ID**: photoluminescence | polypropylene | ultraviolet photons | photoluminescence |
- **irradiation atmosphere** | O/sub 2/ irradiation atmosphere | vacuum (Portion omitted) |
- **CC**: A8160J Surface treatment and degradation of polymers and plastics_ | A6140K (Portion omitted) CI- ArF bin Ar bin F bin (Elements 2)^O2 el O el (Elements 1) |
- **NI**: electron volt energy 3.0E+00 eV^electron volt energy 4.0E+00 eV |
- **CG**: Copyright 2001, IEE | |
Dialog Record Creation and Search

DATABASE PRODUCER

produces records
searchable fields
identifies fields

DIALOG

field labels
word position
noted

Basic index created
Additional index created to
add value

Inverted File
Term -> Accession No.,
Field, Position

SEARCH

Linear File
Accession No.
-> Record
Search Engines

• **Web search engines**: probabilistic information retrieval methods. Count term occurrence and rank on terms. Many variations on probabilistic information retrievals.

• **Dialog**: inverted index.
  - Searchable fields are parsed (separated and extracted into an alphabetical list so they can be searched), Records divided into fields and position noted.
  - Stop-words excluded. Occurrence of term stored in index by accession number, field and position.
  - Both basic (title, abstract, descriptor) and additional indexes (author, journal, CAS RN’s etc) created.
Search Engines

- Dialog and Database aggregator systems
- Basic and Additional Indexes are created.
- Linear and Inverted File.
- Database structure in Bluesheets
- Truncation and internal wild cards. WOM?N
- Internet search engines
- Spiders or crawlers
- Page Rank relevancy
- No standardization of terms
- Truncation variable.
Search Process on Dialog

BEGIN

DB_1, DB_2
... DB_60

RECORD > INDEX
BASIC, ADD'L

SEARCH (240 char)

ADD

DB #1, DB #2
... DB #60

REPEAT

PR S#/N/1-N
50,000 limit

MOUNT DB

REPORT S#/F1D1, FLD2, FLD3/1-N

ORDER
100 records MAX

KEEP S0

FROM DB No

SORT
S#/ALL/FLDS

DATA PROCESSES

READ

DISPLAY S#/n/1-p

DISPLAY S#/n/1-p

TYPE S#/n/1-p

SET S (400 SET S max)

KEEP S#/nnn

ADD'L

BASIC

KEEP S0
Deep Web

- Deep Web: To access data on the “deep web” (data which is in non-HTML format image, video, audio, Macro Media Flash, CGI scripts, archives or proprietary) data must be retrieved using directed query engines.

Figure 1. Search Engines: Dragging a Net Across the Web's Surface

Search techniques

- Scientists use of the Internet is ubiquitous.
- Scientist’s (end-user) searches generally use keywords and rarely use the full capabilities of the search engines.
- Scientists know the subject matter.
- Information professionals are familiar with database structure and effective searching methods.
- Information professionals are generalists and do not always know the discipline.
# Major Search Engines (Web)

<table>
<thead>
<tr>
<th>Engine</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOL Search</td>
<td><a href="http://search.aol.com/">http://search.aol.com/</a></td>
</tr>
<tr>
<td>Aewi</td>
<td><a href="http://www.aewi.com/">http://www.aewi.com/</a></td>
</tr>
<tr>
<td>Altavista</td>
<td><a href="http://www.altavista.com/">http://www.altavista.com/</a></td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td><a href="http://www.askjeeves.com/">http://www.askjeeves.com/</a></td>
</tr>
<tr>
<td>Diabolos</td>
<td><a href="http://www.diabolos.com/">http://www.diabolos.com/</a></td>
</tr>
<tr>
<td>Direct Hit</td>
<td><a href="http://www.directhit.com/">http://www.directhit.com/</a></td>
</tr>
<tr>
<td>DisInformation</td>
<td><a href="http://www.disinfo.com/">http://www.disinfo.com/</a></td>
</tr>
<tr>
<td>Entireweb.com</td>
<td><a href="http://www.entireweb.com/">http://www.entireweb.com/</a></td>
</tr>
<tr>
<td>EuroSeek</td>
<td><a href="http://www.euroseek.com/page?ilang=en">http://www.euroseek.com/page?ilang=en</a></td>
</tr>
<tr>
<td>Excite</td>
<td><a href="http://www.excite.com">http://www.excite.com</a></td>
</tr>
<tr>
<td>Fastsearch</td>
<td><a href="http://www.alltheweb.com/">http://www.alltheweb.com/</a></td>
</tr>
<tr>
<td>Google</td>
<td><a href="http://www.google.com">http://www.google.com</a></td>
</tr>
<tr>
<td>Hotbot</td>
<td><a href="http://hotbot.lycos.com/">http://hotbot.lycos.com/</a></td>
</tr>
<tr>
<td>Iwon</td>
<td><a href="http://hotbot.lycos.com/">http://hotbot.lycos.com/</a></td>
</tr>
<tr>
<td>Inktomi</td>
<td><a href="http://www.inktomi.com/search">http://www.inktomi.com/search</a> engine</td>
</tr>
<tr>
<td>LookSmart</td>
<td><a href="http://www.looksmart.com/">http://www.looksmart.com/</a></td>
</tr>
<tr>
<td>Lycos</td>
<td><a href="http://www.lycos.com/">http://www.lycos.com/</a></td>
</tr>
<tr>
<td>MSN Search</td>
<td><a href="http://search.msn.com/">http://search.msn.com/</a></td>
</tr>
<tr>
<td>Netscape Search</td>
<td><a href="http://search.netscape.com/">http://search.netscape.com/</a></td>
</tr>
<tr>
<td>Northern Light</td>
<td><a href="http://www.northernlight.com/">http://www.northernlight.com/</a></td>
</tr>
<tr>
<td>Open Directory</td>
<td><a href="http://dmoz.org/">http://dmoz.org/</a></td>
</tr>
<tr>
<td>Qango.com</td>
<td><a href="http://www.qango.com/">http://www.qango.com/</a></td>
</tr>
<tr>
<td>RealNames</td>
<td><a href="http://www.realnames.com/Eng/Eng_Corporate_RealNamesHomepage.asp">http://www.realnames.com/Eng/Eng_Corporate_RealNamesHomepage.asp</a></td>
</tr>
<tr>
<td>Teoma Technologies</td>
<td><a href="http://www.teoma.com/">http://www.teoma.com/</a></td>
</tr>
<tr>
<td>Web Crawler</td>
<td><a href="http://www.webcrawler.com/">http://www.webcrawler.com/</a></td>
</tr>
<tr>
<td>WebTop</td>
<td><a href="http://www.webtop.com">http://www.webtop.com</a></td>
</tr>
<tr>
<td>Yahoo</td>
<td><a href="http://www.yahoo.com/">http://www.yahoo.com/</a></td>
</tr>
<tr>
<td>4anything.com</td>
<td><a href="http://www.4anything.com/">http://www.4anything.com/</a></td>
</tr>
</tbody>
</table>
# Metacrawlers

<table>
<thead>
<tr>
<th>Dogpile</th>
<th>Ixquick</th>
<th>MetaCrawler</th>
<th>Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vivisimo</td>
<td>QbSearch</td>
<td>ProFusion</td>
<td>InfoGrid</td>
</tr>
<tr>
<td>SurfWax.com</td>
<td>TeRespondo</td>
<td>Mama</td>
<td></td>
</tr>
</tbody>
</table>

**Major metacrawlers** (from [http://searchenginewatch.com/links/Major_Search_Engines/](http://searchenginewatch.com/links/Major_Search_Engines/))
How “Metacrawlers” work

- Agent computer program sends a search to multiple search engines at the same time.

- Results ranked based on search term frequency.

- Evaluate search results according to authority, accuracy, currency, objectivity and coverage.
USER >>>>>> QUERY
USER <<<<<< RESULTS

QUERY TRANSLATED by METACRAWLER

Lycos  Yahoo  Open Text  InfoSeek  Inktomi
Interfaces and Database Structure

- Command Line vs. Graphical user interface
- Non-graphical vs. Graphical, icons
- Remember commands vs. Icons
- Fast, simple vs. Easy to learn, little intuitive
- C:>. DOS vs. Windows, Mac
- telnet, pine vs. Wintelnet, Eudora
Search Strategies

- Web based searching: keyword search techniques and probabilistic retrieval.
- Not uniform throughout the web.
- Must use results from a one-pass strategy

- Online Information Retrieval Services:
  - Boolean search: AND, OR NOT.
  - Uniform search mechanism.
  - Can repeat and refine search by repeating it on retrieved sets and adding databases.
Dialog Search Strategy

BEGIN
MOUNT DB DB_1, DB_2 ...
RECORD > INDEX BASIC, ADD'L

SEARCH (240 char)
ADD DB_No2, DB_No3 etc.

KEEP S0
MOUNT DB
REPEAT

SET S#/n/1-p
SORT S# /ALL/FLDS
DISPLAY S#/n/1-p
REPORT S#/ FLD1, FLD2, FLD3/1-N
PRINT S#/n/1-p
TYPE S#/n/1-p

ORDER 100 records MAX
KEEP S#/nnn FROM DB_No.
RD, ID
DISPLAY S#/n/1-p
DESCRIPTION
S#1
S#2
S#3
DIALOG SEARCH STRATEGY

S (Term_1? AND Term_2)/ti,de,ab <S1>
S (Term_3 OR Term_4?)/ti,de,ab    <S2>
S Term_5 AND Term_6/ti,de,ab      <S3>
S (S1 AND S2) NOT S3                 <S4>

KEEP S#/N,O,P

SEARCH RESULTS

RANK No. ITEMS TERM
1 nn term_1
2 oo term_2
... ... ....
N qq term_N

FOLLOW MENU
PROMPT
select terms
1,2,5, N

CREATE SEARCHES w. fields. MAP

Trend analysis of terms: RANK

RANK FIELDS

KEEP S#/N,O,P

MAP FIELDS T

BEGIN New_Files;
EXS

FREE TEMP ;
EXS;
S S# AND S#;
SAV E TEMP_1
Text based Information Retrieval

- Searched sets can be reused by
- 1) isolating terms
- 2) isolating records
- 3) mapping fields and changing prefixes
- Add additional databases and repeat the search in these new added databases
Chemical Search Techniques

1) Structure -> linear search string
- ROSDAL (Representation Of Structure Diagram Arranged Linearily)
- SMILES (Simplified Molecular Input Line Input Line Entry Specification)
- MARKUSH (generic structures for patents)
- Effective search uses connection tables.
- CHEMICAL NAMES or CAS RN’s can be used
Chemical Search Techniques
Central File in the Sciences

- BIOSIS
- SCISEARCH
- AGRICOLA
- CAB ABSTRACTS
- ENVIROLINE
- FDC-REPORTS
- PHIND
- ADIS R&D
- PHARMAPROJECTS
- EMBASE
- MEDLINE
- TOXFILE
- EXTRAMED
- HEALTHSTAR
- NEJM
- LANCET
- AMA Journals
- INSPEC
- EI COMPENDEX
- CHEMENG
- RAPRA (Rubber/Polymer)
- CERAMICS
- PAPERCHEM
- PIRA

**CHEMNAME**
- B 398
- CAMEX
- RNLOOKUP
- CHEMLIT
- CA SEARCH
- FOODSCI
- ENG

**BIOSCI**
- PHARM
- MEDICINE
- MEDTEXT
- STRUCTURE SEARCH

**OneSearch**
- TD001 (MAP SYRN T)
- USED IN FILES WITH RN
- TD002 (MAP SY T)
- USED IN FILES WITHOUT RN

**Prior art**
- S 51/PAT
- MAP PN T
- SAVE TEMP
- PATENT FILES
MAP and KEEP command

- KEEP and MAP commands allow isolation of defined records, excision of defined file prefixes and creation of a new search.

- The new search statements can be executed in a new file or files.
Use of KEEP, MAP commands

Prior Art
- Identify chemical name
- Search Science Files

Identifier
- Isolate Patent Record
- Initial Patent
- Cited Patent
- Identify Cited Patent Records

KEEP S#/record_no (isolate relevant record)
S S0 (create a new search statement)

MAP PN T (create temporary file with Patent Number)

B 351 Derwent World Patents Index
exs (Identify parent patent)
MAP PN T

B 342 (Derwent Patent Citation Index)
exs
MAP RXAX T (Identify Citing Patents [related patents])

B 351 (Derwent World Patents Index)
exs
Type out titles of relevant records (use format 26)

E NA=chemical name
MAP SYRN T

B Files_of_interest
S terms
Type out titles and identify relevant records
S s1 from 399 (patent file)
S S1/pat (isolate patents)
MAPping Allows Extraction of Fields between Databases

- This can be done for Patent Databases.
- Much industrial data is present in patents.
- This enables coordinated searches to be done between databases to isolate a more complete retrieval set.
Inter-relationship of MAPped Fields for Patent Files

- **MAP AN T** or **MAP ANPR T** followed by **EXS** in the same file. Extracts application numbers and iterates search the same file for additional members of a family.
- **MAP PNCT T** or **MAP CTPN T** or **MAP PNCT=** followed by **EXS** will extract patent numbers and search for the cited patents.
- If MAP not available, **USE RANK** and save all entries in RANK List.

**REFERENCES TO NONPATENT LITERATURE**
- CX = Chemical Abstracts Number
- RF = Cited Reference

**FIELD NAME**
- AN = PATENT APPLICATION NUMBER
- CN = CHEMICAL NAME
- CO = COMPANY NAME
- CT = CITING PATENT
- IC = INTERNATIONAL CLASSIFICATION
- PN = PATENT NUMBER
- RN = CAS REGISTRY NUMBER
- SY = SYNONYM

**IDPAT**
- manage overlapping patent data by organizing of patent records into family-related groups and records with duplicates identified.
File Linkages for Pharmaceutical Files

TC = PROUSE THERAPEUTIC CLASS

File 452 Prous Drug Data Report
MAP
AAT
ST=Phase I, II, III
MAP CX T
MAP SS T
File 453 Prous Drugs of the Future
MAP RN T
EXS
File 375 Derwent Drug Registry
MAP RN T
EXS
File 377 Derwent Drug File
MAP RN T
EXS

File 128 Pharmaprojects
MAP RN T
EXS
ST=COUNTRY
File 398 Chemname
MAP RN T
EXS
Phase I, II, III
File 107,907 ADIS R&D Insight
MAP RN T
EXS
File 229 Drug Information Full-text
MAP RN T
EXS

File 453 Prous Drugs of the Future
MAP RN T
EXS
ST=Phase I, II, III
AC=EPHmra ATC CODES
AC=WHO ATC CODES
SC=THERAPEUTIC VALUE
0-44, 43-59, 60-79, 80-100
ES=CLINICAL EVALUATION SCORE
0-50, 50-70, 71-85, 86-100
STUDY QUALITY

LINKAGES BETWEEN PHARMACEUTICAL FILES ON DIALOG.
Summary of MAPping

- MAP allows data and field identifiers to be isolated from target files and to create a new search statement which can be executed in new files.

- This allows interactive searches on multiple databases. Iterative searches are difficult to accomplish in web-based searches.
Multivariate Analysis

- Textual data of term frequencies are extracted from documents (abstracts).
- Statistical techniques used in document association: Cluster and factor analysis.
- Principal components analysis (PCA) and Factor Analysis (DA) are applied to group documents.
Summary

- Text based information offers retrieval and data interchange through MAP commands and synonym isolation.

- Chemical searches are effectively done using well designed textual searches and the CAS Registry Number.

- Textual information retrieval is amenable to multivariate analysis.
Thank you

THE END