It’s about nothing.
The electronic flood

- Drinking from a fire hydrant
Language fundamentally dynamic

- Role for learning
Finding out about - a cognitive activity

- Organisms - complex environment
- Language users - Symbols
- Academics - Language archives
- The FOA process
- Search Engine Schematic
Organisms - complex environment

- Actively seeking ...
- ... useful information
Language users - Symbols

- Oral first
- Written language
  - External memory
  - Conventional
- Used to guide FOA search
Academics - Language archives

- Electronic artifacts
- Libraries
- Textbooks
- Language becomes object of search, too
The FOA process

1. The asking of a question
2. The construction of an answer
3. Assessment of the answer
4. Query sessions
The asking of a question

- The “user” has an “information need”
- Ill-defined, cognitive state
The construction of an answer

- Forming a good answer is hard!
- *Assume*: answer = {predefined text passages}
Assessment of the answer

- Closing the ask/answered loop
Query sessions

- There are no ‘magic bullet’ queries
Search Engine Schematic

Information need

Query

Retrieved Documents

Relevance Feedback

Match

Information retrieval system

Indices

Documents
KW - Overview

• Elements of the query language
• Characterizations of documents’ subject matter
  • "The documents' description language" [vR 22]
• Vocabulary
• Topical scope
• Document descriptors
Words

- Morphology
- Stemming algs.
- (Counter) examples
  - NEUTRAL/NEUTRON
  - ORIGINAL/ORIGINALITY
Phrases

- Syntax
- Noun phrases
- cf. German
  - Unterhaltungsindustrie (show business)
  - Geschwindigkeitsbeschränkung (speed limit)
Query syntax

- Expressions of an “information need”
- In some language
- Query sessions
Expressions of an “information need”

- Cognitive state; hidden
- Paradoxical: characterizing ignorance
In some language

- Expression; not hidden
- Boolean connectives
  - BUT Not
- Proximity
Query sessions

IRS

Query

Q₀

Q₁

Q₂

Ret₁

Ret₂

Retₙ
“Passage of free text”

- Character set
- Natural, written language
- Genre
Arbitrary size

- Normalize lengths
  - A1: All documents have equal 'aboutness'.
- Focus on paragraphs
  - A2a. The smallest unit of text is the paragraph.
  - A2b. All manner of longer documents can be constructed out of these basic atoms.
## Structured aspects of docs

- hierarchical containment composing paragraphs into sub-sections, sections, chapters, etc.
- footnotes, embellishing the primary theme;
- bibliographic citations to other, previous publications;
- references to other sections of the same document; especially
- pedagogical prerequisite relationships ensuring that conceptual foundations are established prior to subsequent discussion; etc.
Often (database) attributes associated, too

- Publication facts
- Authorship
- Authority lists: normalizing
POLITICAL FIDELITY

- PRE- MONICA!
Obsolete concert schedule

[Image of an obsolete concert schedule with a note saying to ignore the calendar as the dates are 3 years old.]

Text only one example of a larger class of data

- Not possible to fully enumerate all facts
- Eg, Kremlinology
- Also images, music, speech, videos, ...
- Some FOA techniques generalize to these media, too
Indexing

- Mathematical view
- Manual/automatic
Mathematical view

- Index: A map from documents to the keywords they are about
  \[ \text{Index} : \; \text{doc}_i \xrightarrow{\text{about}} \{ \text{kw}_j \} \]

- Inverted: From keywords to the documents they describe
  \[ \text{Index}^{-1} : \; \text{kw}_i \xrightarrow{\text{describes}} \{ \text{doc}_j \} \]
Manual/automatic

- If manual, By whom?
- Derived, or independent
- A role for machine learning!

Ancillary structure

Free text

Manual encoding effort

Editorial enhancement

Encyclopedia Britannica
WESTLAW
A role for machine learning!

- Ancillary structure
- Editorial enhancement
  - *Encyclopedia Britannica*
  - *WESTLAW*
- Free text
- Learning
  - Manual index
  - Automatic index
- Manual encoding effort
<table>
<thead>
<tr>
<th>SYSTEM PROVIDES:</th>
<th>Information retrieval</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference to data</td>
<td></td>
<td>data</td>
</tr>
<tr>
<td>USER'S QUERY:</td>
<td>general</td>
<td>specific</td>
</tr>
<tr>
<td>RETRIEVAL:</td>
<td>probabilistic</td>
<td>deterministic</td>
</tr>
<tr>
<td>SUCCESS CRITERIA:</td>
<td>utility</td>
<td>(correctness)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>efficiency,</td>
</tr>
</tbody>
</table>
|                         |                       | user-friendliness,
Sample query

• “What is the best SCSI disk drive to buy?”

• Database relation:

<table>
<thead>
<tr>
<th>Model number</th>
<th>Manufacturer</th>
<th>Vendor</th>
<th>Size (GB)</th>
<th>Interface</th>
<th>Price ($)</th>
<th>Speed (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>Seag</td>
<td>J&amp;R</td>
<td>2.4</td>
<td>SCSI</td>
<td>162</td>
<td>12</td>
</tr>
<tr>
<td>123</td>
<td>Seag</td>
<td>Fry</td>
<td>2.4</td>
<td>SCSI</td>
<td>159</td>
<td>12</td>
</tr>
<tr>
<td>456</td>
<td>Metrop</td>
<td>A&amp;B</td>
<td>2.5</td>
<td>IDE</td>
<td>$0</td>
<td>12.5</td>
</tr>
<tr>
<td>789</td>
<td>Seag</td>
<td>J&amp;R</td>
<td>1.5</td>
<td>EIDE</td>
<td>121</td>
<td>10.5</td>
</tr>
</tbody>
</table>

...
## UseNet search

Matches 1-20 of 726 for search: best SCSI disk

<table>
<thead>
<tr>
<th>Date</th>
<th>Scr</th>
<th>Subject</th>
<th>Newsgroup</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 97/08/29</td>
<td>062</td>
<td>Re: IDE or SCSI?</td>
<td>comp.os.ms-windows</td>
<td>hima@ur</td>
</tr>
<tr>
<td>2. 97/08/04</td>
<td>053</td>
<td>Re: switching from IDE to SCSI</td>
<td>comp.unix.unixware</td>
<td>arthur@ur</td>
</tr>
<tr>
<td>3. 97/08/28</td>
<td>051</td>
<td>Re: Sun SCSI and Linux</td>
<td>comp.os.linux.hardw</td>
<td>tallpav</td>
</tr>
<tr>
<td>4. 97/08/03</td>
<td>050</td>
<td>Re: switching from IDE to SCSI</td>
<td>comp.unix.unixware</td>
<td>Larry F</td>
</tr>
<tr>
<td>5. 97/08/02</td>
<td>050</td>
<td>Replace SCSI Disk on R3.</td>
<td>comp.unix.sco.misc</td>
<td>ericl@r</td>
</tr>
<tr>
<td>6. 97/08/21</td>
<td>049</td>
<td>FS: 466-100Hz multimedia</td>
<td>misc.forsale.comput</td>
<td>tjioe@r</td>
</tr>
<tr>
<td>7. 97/08/11</td>
<td>049</td>
<td>SCSI problems</td>
<td>comp.os.linux.hardw</td>
<td>//@ax 2</td>
</tr>
<tr>
<td>8. 97/08/03</td>
<td>049</td>
<td>Re: Which is Best: More</td>
<td>comp.os.linux.hardw</td>
<td>o_r_c @</td>
</tr>
<tr>
<td>9. 97/08/01</td>
<td>049</td>
<td>Best Scsi CDR for $300-$</td>
<td>alt.cd-rom</td>
<td>Scott</td>
</tr>
<tr>
<td>10. 97/08/16</td>
<td>048</td>
<td>Re: IDE increases Interr</td>
<td>microsoft.public.wi</td>
<td>&quot;Nolan</td>
</tr>
<tr>
<td>11. 97/08/16</td>
<td>048</td>
<td>FS: Micropolis 3243WAV dr</td>
<td>rec.video.marketpla</td>
<td>John Cc</td>
</tr>
<tr>
<td>12. 97/08/09</td>
<td>048</td>
<td>CDR Server --- Audio CD</td>
<td>bk.forsale</td>
<td>Mattgor</td>
</tr>
<tr>
<td>13. 97/09/01</td>
<td>047</td>
<td>comp.sys.apple2 Freq#8/1</td>
<td>comp.sys.apple2</td>
<td>nathan@</td>
</tr>
<tr>
<td>14. 97/08/31</td>
<td>047</td>
<td>Re: W95 vs. MacOS Micros</td>
<td>comp.sys.mac.advoca</td>
<td>dfield@</td>
</tr>
<tr>
<td>15. 97/08/27</td>
<td>047</td>
<td>mirco configuration utili</td>
<td>rec.video.desktop</td>
<td>Edy Gas</td>
</tr>
<tr>
<td>16. 97/08/05</td>
<td>047</td>
<td>[FOR SALE] Dual PowerPC</td>
<td>comp.sys.be.misc</td>
<td>&quot;Frank</td>
</tr>
<tr>
<td>17. 97/08/04</td>
<td>047</td>
<td>Re: IBM OEM 0664-CSH and</td>
<td>comp.periph.s.scsi</td>
<td>Chuck T</td>
</tr>
<tr>
<td>18. 97/08/01</td>
<td>047</td>
<td>error 1 kernel32.dll alt</td>
<td>no.pc</td>
<td>Jorunn</td>
</tr>
<tr>
<td>19. 97/08/12</td>
<td>046</td>
<td>SUN SPARC 20 FOR SALE.</td>
<td>comp.sys.sun.hardwa</td>
<td>power@q</td>
</tr>
<tr>
<td>20. 97/07/26</td>
<td>046</td>
<td>Re: Help: Replacing pri</td>
<td>comp.os.ms-windows</td>
<td>Christo</td>
</tr>
</tbody>
</table>
Typical answer

Subject: Re: IDE or SCSI?
From: helpful@urban.or.jp (me @ my humble abode)
Date: 1997/08/29
Message-Id: <3406f93a.50446926@nnrp.gol.com>
Newsgroups: comp.os.ms-windows.nt.setup.hardware,comp.windows.nt.misc

On Thu, 28 Aug 1997 23:10:03 GMT, Michael Query
<query@dpi.qld.gov.au> wrote:

>My question is, should I get another 2 Gb SCSI disk for putting the
>OS (NT 4.0 WS), software, etc on, or should I get an IDE disk
>for this?

Having played around with different configs for a while, I'd say go SCSI. I'd do that even if I had to get a second SCSI controller.

(You'll "hear" a lot of people arguing that IDE is good enough, but if you are after overall improved performance SCSI is best.)

my 2Y.
How are we doing?

- Single user’s perspective
- Search engine builder’s perspective
- Summary
Single user’s perspective

- RelFbk!
Comparison against ‘relevant’

- High precision
- High recall

Retrieved

Relevant
Precision & Recall

\[
\text{Precision} \equiv \frac{|\text{Ret} \cap \text{Rel}|}{|\text{Ret}|}
\]

\[
\text{Recall} \equiv \frac{|\text{Ret} \cap \text{Rel}|}{|\text{Rel}|}
\]
Summary

- FOA is a natural activity
- It involves language as part of a three part dialog between question-asker and answer.
- Computer science can provide many tools to make the process more efficient
- Relevance feedback provides key information from which we can learn a great deal
- Evaluating the performance of our tools is the only way to make progress