

Information from Text: Overview of Background and Opportunities*

Steven O. Kimbrough
University of Pennsylvania
565 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, PA 19104-6340
kimbrough@wharton.upenn.edu
215-898-5133

May 15, 2003

*File: open-tech-foils.pdf.

Context & Framing

- Aim: Overview problems and opportunities in extracting useful information from text.
- Open, 'technical' briefing. Style: graduate seminar.
- Some items here depend upon proprietary methods and/or results, which will not be revealed.
- Key distinction:
structured data vs. unstructured data vs. semi-structured data

The Product Placing (P^2) Problem

Consider the prototypical case of a firm having the rights to C, an intellectual property (IP). Think of C as a component product, rather than an end product, E. C is neither shoes, nor ships, nor sealing wax. Instead, C may be a valuable component (e.g., substance or process) in making shoes, or ships, or sealing wax, etc. The *product placing* (P^2) problem has these main aspects:

1. For which end products, the Es, is C a promising component?
2. Who, including possible partners, is well-positioned to commercialize the Es, using C?

In addition, consider an end product, E, or a problem, P. Think of P as a recognized shortcoming in a recognized end product. P is too much labor content, too little resistance to sunlight, unwanted reliance on nonrenewable resources, and so on. (Cost reduction is always and available P.) Thus, the product placing problem also has these main aspects:

3. For a given end product, E, find previously unrecognized components, Cs, with promise of improving E.
4. For a given problem, P, find previously unrecognized end products, E, using component products C, that provide an improved solution to the problem.
5. Who, including possible partners, is well-positioned to commercialize such improvements to product E and solutions to problem P?

We Envision: P² Support System

- Thoroughgoing support for P² problems; DSS concept
- Built upon Sizatola concepts and experience to date
- Add additional algorithms for retrieval, visualization, etc.
- Acquire and exploit additional data and document collections

Fundamentals now. . .

Exploiting Weak Structure: Text Mining

PageRank™

1. Underlies Google
2. How does it work?
3. Lesson #1: There are clever ways to extract information automatically from weakly-structured sources.
4. Lesson #2: (later)

PageRank Explained

PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."

Important, high-quality sites receive a higher PageRank, which Google remembers each time it conducts a search. Of course, important pages mean nothing to you if they don't match your query. So, Google combines PageRank with sophisticated text-matching techniques to find pages that are both important and relevant to your search. Google goes far beyond the number of times a term appears on a page and examines all aspects of the page's content (and the content of the pages linking to it) to determine if it's a good match for your query.

[From: <http://www.google.com/technology/>]

Exploiting Weak Structure: The Plato Problem

Imagine that you have access to a large document collection and are interested in obtaining information about Plato, the ancient Greek philosopher. You successfully retrieve all documents containing the word 'Plato'. When you examine these documents you find what you expected: many of them are indeed about the Greek philosopher; others are about various commercial products trading on the philosopher's name, but having nothing to do with him.

You are then surprised when a friend presents you with a completely different set of documents, also retrieved from the same large document collection. Nearly all of these new documents are very much about Plato the philosopher, yet none of them mention 'Plato' (or any close term, e.g., 'Platon' a foreign spelling).

How can this be? Your friend explains. 'Plato' was not the philosopher's original name. It was a nickname given to Aristocles by his wrestling coach ('personal trainer' in today's language) and it means 'chubby' or 'chubs'. Somehow it stuck. Many authors, however, have written about the philosopher using only his original name. These writings would not be retrieved by a keyword search using 'Plato'.

The Plato Problem is the problem of designing a computer system that helps in finding relevant and useful documents that do not contain the search term(s) you have in mind. The scope of this problem extends far beyond the world of classical scholarship. In fact, except in relatively trivial cases, it appears much more often than not.

From the STAIRS study (Blair & Maron)

Sometimes we followed a trail of linguistic creativity through the database. In searching for documents discussing “trap correction” (one of the key phrases), we discovered that relevant, unretrieved documents had discussed the same issue but referred to it as the “wire warp.” Continuing our search, we found that in still other documents trap correction was referred to in a third and novel way: the “shunt correction system.” Finally, we discovered the inventor of this system was a man named “Coxwell” which directed us to some documents he had authored, only he referred to the system as the “Roman circle method.” Using the Roman circle method in a query directed us to still more relevant but unretrieved documents, but this was not the end either. Further searching revealed that the system had been tested in another city, and all documents germane to those tests referred to the system as the “air truck.” At this point the search ended, having consumed over an entire 40-hour week of on-line searching, but there is no reason to believe that we had reached the end of the trail; we simply ran out of time.

The Plato Problem: Computational response?

- Intuition: Relevant documents with the term you search on look rather like relevant documents without the search term.
- Find a profile of terms for the relevant documents having the search term, then find similar documents without the term.
- Such algorithms can rank collections by relevance.
- Expensive. Good performance.

The Plato Problem: Computational response (con't.)

- Concept:

For any given query based on search term(s) t there are conceptually three kinds of documents:

1. Relevant, containing t
2. Relevant, not containing t
3. Not relevant

N.B. We'd like a ranking by relevance, too.

- Intuition/idea:

Profile the documents in class 1, then use the profile to separate class 2 from class 3.

How to Do This?

- Many ways, mostly not explored.
- One (simple) way, developed and explored with positive results in Kimbrough's lab: DCB representation. See "On Relevance and Two Aspects of the Organizational Memory Problem" (orgmems5.pdf).
- DCB representation versus algorithm(s).

Here: representation. Proprietary: fast algorithms for calculation.

DCB Representation: Step 1, the K Matrix

Document representation: The indexing is done by determining, for every document and every keyword in an appropriately thorough list, whether a document contains a keyword. Having done this, we in effect have a matrix, called K , whose entries are all 1s and 0s, whose rows correspond to keywords, and whose columns correspond to documents. A particular K might look like this, where: rows = keyterms and columns = documents.

$$K = \begin{pmatrix} 1 & \dots & 0 & 1 \\ \vdots & k_{i,j} & \vdots & \vdots \\ 1 & \dots & 1 & 0 \end{pmatrix} \quad (1)$$

Element $k_{i,j}$ is 1 if keyword i occurs (at least once) in document j ; otherwise it is 0.

DCB Representation: Example

$$K = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix} \quad (2)$$

Taking a simple example, suppose we have six keywords and six documents, with the following K matrix. (The fact that K is square in this example is not significant. Our remarks apply no matter what shape K has.) Consider: relevance to keyterm 1 (first row). How should the documents (columns) be ranked? Note: The Plato problem.

DCB Representation: Step 2, the L Matrix

$$L \stackrel{\text{def}}{=} K \cdot K^T = \begin{pmatrix} 4 & 3 & 1 & 2 & 1 & 2 \\ 3 & 5 & 2 & 3 & 1 & 2 \\ 1 & 2 & 2 & 2 & 0 & 2 \\ 2 & 3 & 2 & 3 & 0 & 2 \\ 1 & 1 & 0 & 0 & 2 & 1 \\ 2 & 2 & 2 & 2 & 1 & 3 \end{pmatrix} \quad (3)$$

The interpretation of L is straightforward. L is always a square, symmetric matrix with number of rows (and columns) equal to the number of keywords, i.e., the number of rows in K . The diagonal elements of L indicate the number of documents containing the corresponding keyword. In our example, $l_{2,2} = 5$, indicating that 5 documents contain keyword ii , which may be verified by examining K . The off-diagonal elements, the $l_{i,j}$ s, indicate the number of documents containing *both* keyword i and keyword j .

DCB Representation: Step 3, the M Matrix

$$M \stackrel{\text{def}}{=} K \cdot K^T \cdot K = L \cdot K = \begin{pmatrix} 12 & 8 & 9 & 4 & 7 & 7 \\ 15 & 12 & 11 & 6 & 6 & 8 \\ 9 & 8 & 5 & 2 & 3 & 3 \\ 12 & 10 & 8 & 3 & 4 & 5 \\ 3 & 2 & 2 & 3 & 4 & 2 \\ 11 & 9 & 6 & 3 & 6 & 4 \end{pmatrix} \quad (4)$$

Looking at M , we see that the DCB algorithm has produced a relevancy ranking of all the documents, and in doing so uses a form of associative retrieval. Step 3 of the DCB algorithm produces the DCB sort: looking at the row of M corresponding to the keyword search term, treat the entries as scoring the relevance of their corresponding documents, with higher numbers indicating greater relevance.

DCB Discussion

1. Matrix multiplication is expensive.

SOK has a proprietary fast algorithm for DCB. Also, for smallish collections, standard methods will do.

2. Does it work? Apparently yes. See the paper, cited earlier. Positive results from initial experiments. Note: A principled solution to the Plato problem.

3. Other methods? Yes, notably LSI, which has recently gone off patent. Also, from the statement of intuition (above) other approaches are designable. Empirical testing is much needed.

The Raynaud Problem

You have a daughter who has been diagnosed with an unpleasant disease called Raynaud's Syndrome. You are told that there is no known effective treatment of the disease, and you quickly verify this by a thorough search of all the on-line medical research papers.

Concerned about your daughter, you wonder if, somehow, a solution has been found, but no one has "made the connection" with Raynaud's Syndrome. The Raynaud Problem is the problem of designing a computer system that will help you in following up on your hunches, specifically here and in general. Can you find a plausible hypothesis for a treatment in the existing literature?

The Raynaud Problem: Computational response?

- Originally, Don Swanson of U. Chicago
- Intuition: there may be undiscovered knowledge implicit in document collections
- Idea: For topic A, find documents on topic B, mentioned by the documents of type A, and find documents on topic C, where the C documents mention the B documents, but the A and C documents don't mention each other. Hypothesize an undiscovered association between topic C and topic A, via topic B.

The Raynaud Problem: Computational response?

- Example: A=documents on Raynaud's. B = documents on blood viscosity. C= documents on fish oil.
- This and other hypotheses confirmed in clinical studies
- Manual discovery subsequently supported with automation and computationally duplicated

Lesson #2 from PageRank:

- Exploit the work of others.

Note: Swanson's work preceded, and was published before, PageRank.

Automating Raynaud

- In “Literature-based discovery by lexical statistics” (*JASIS* 1999) Lindsay and Gordon report on a successful automated duplication of Swanson’s Raynaud findings.
- SOK has devised a method based on use of the DCB matrices. (It has not been tested.)
- A rich field of opportunities.

The Practice Problem

You are a hospital administrator and you would like to have information about how the practices of your emergency room doctors differ. Do they treat similar patients differently? If so, how are the treatments different? Do they see different kinds of patients? and so on. Unfortunately, the only electronic records you have of the patient visits are in document form. There is no database to mine, just a series of emergency room documents.

The Practice Problem is the problem of designing a computer system that will help you find and see meaningful patterns of behavior, based on information extracted from weakly-structured documents. These patterns of behavior can only emerge from the document collection as a whole. The information is not present in any small number of documents. It is implicit in the collection, not explicit in any document.

Garett Dworman: Homer

Patterns in text collections. Use of term-term co-occurrence data extracted from text. Experimental support for efficacy, e.g., with emergency room records in XML.

http://homepage.mac.com/garett_dworman.

http://homepage.mac.com/garett_dworman/homer/HomerER.html

http://homepage.mac.com/garett_dworman/homer/HomerLaughlinArchitecture.html

http://homepage.mac.com/garett_dworman/homer/HomerLaughlinFantasy.html

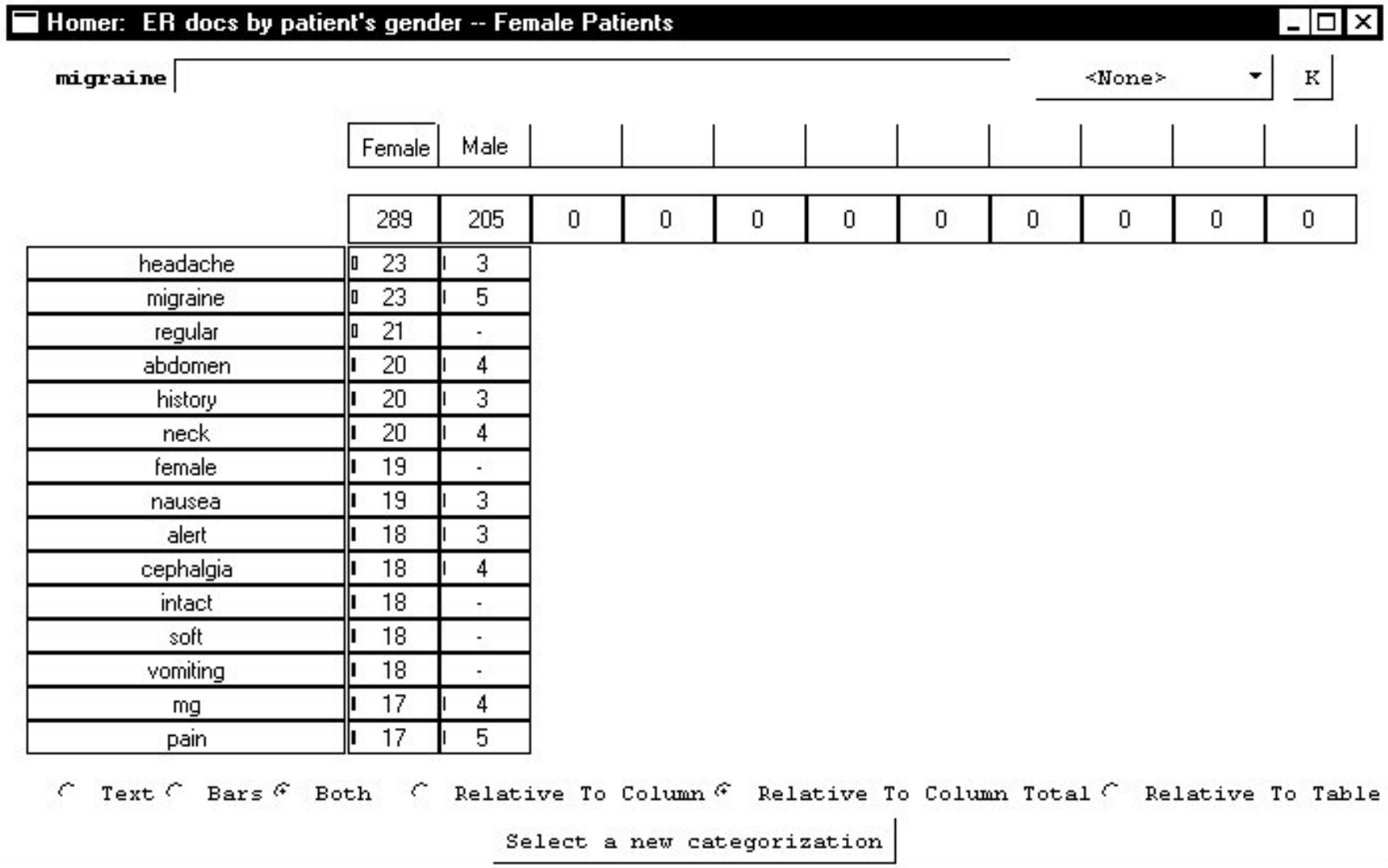


Figure 1: Pattern of Migraine Headaches

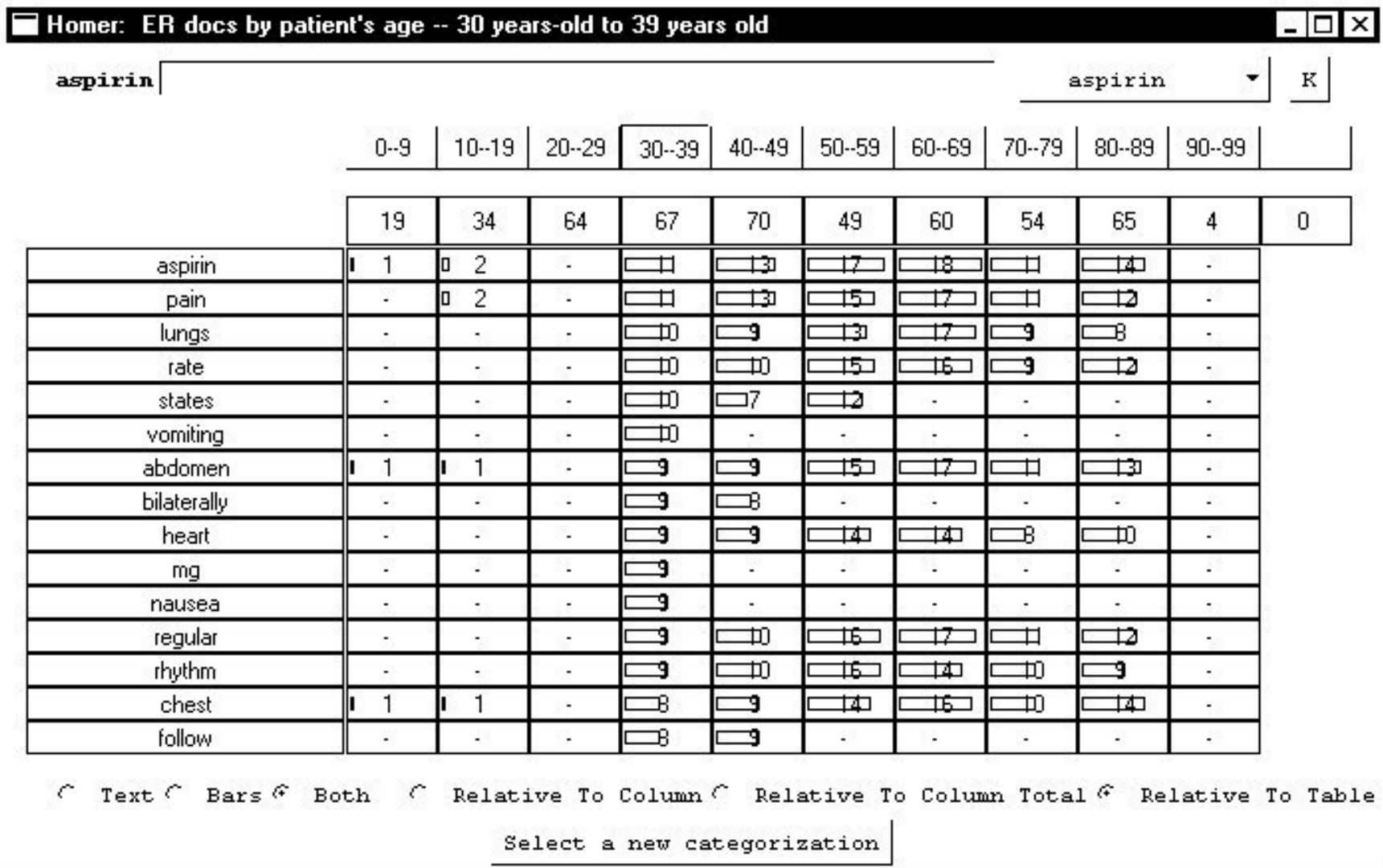


Figure 2: Pattern of Aspirin Use

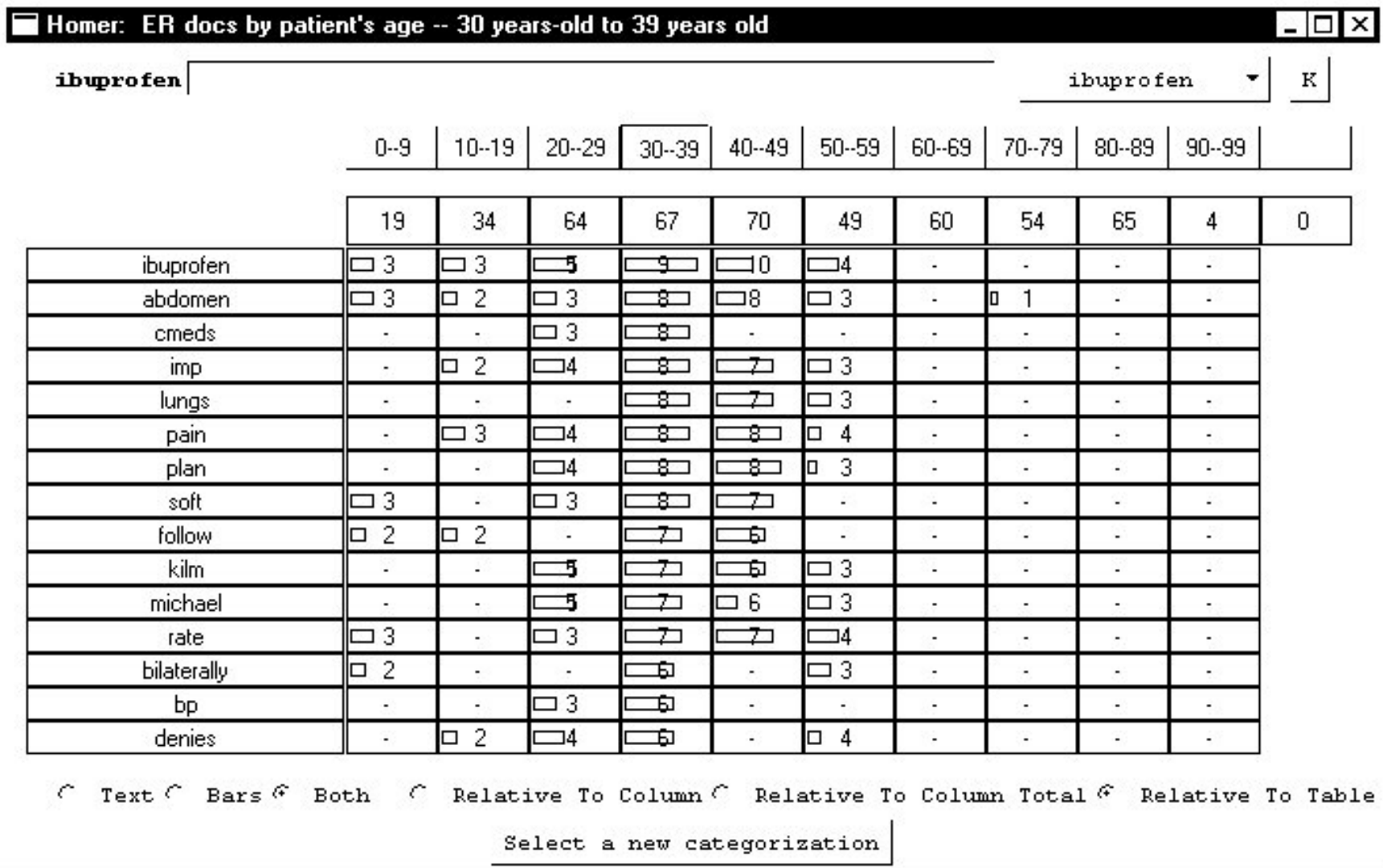


Figure 3: Pattern of Ibuprofen Use

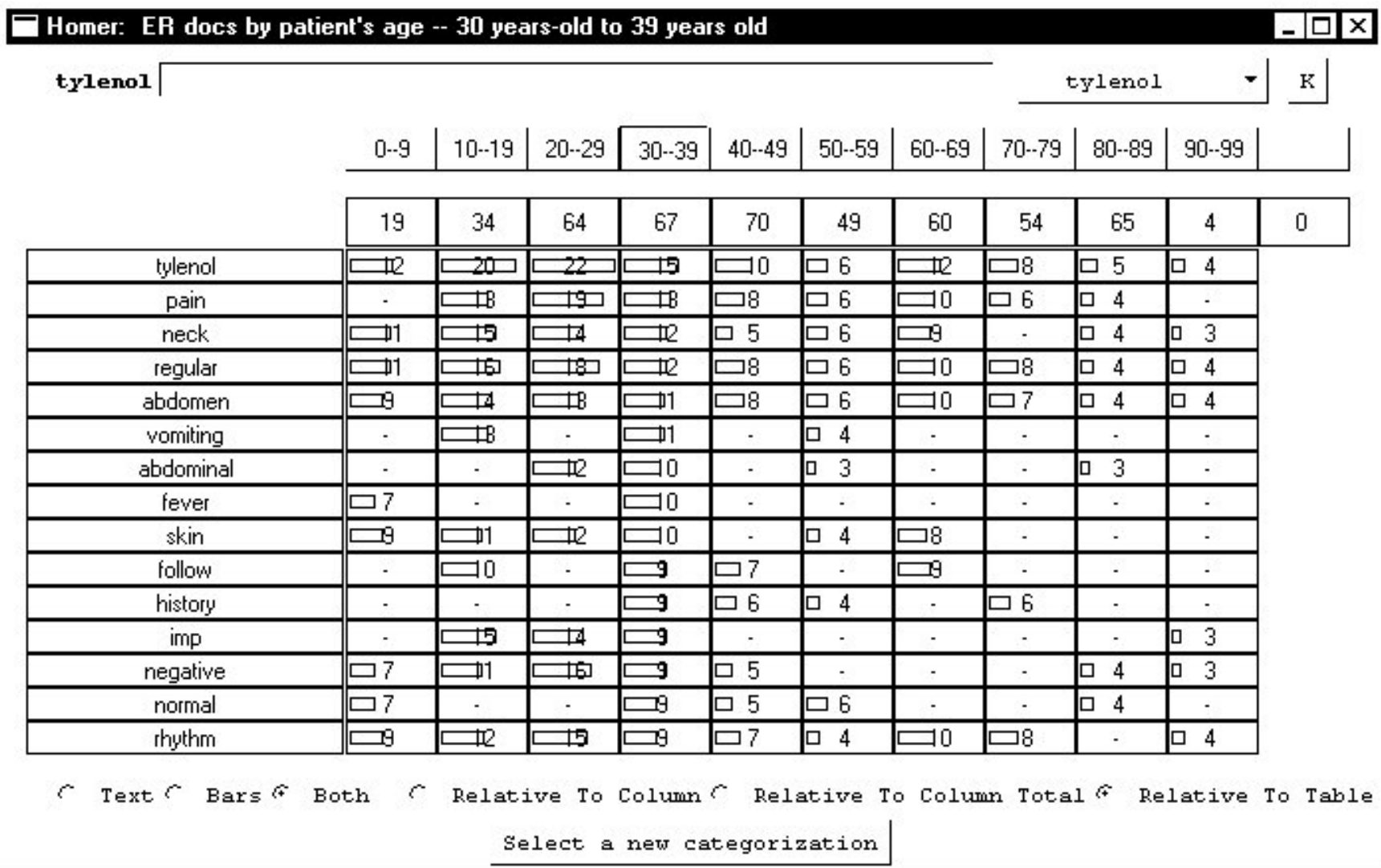


Figure 4: Pattern of Tylenol Use

The Anonymous Problem

You are presented with a text—a sample of speech, a document or a portion of a document—whose authorship is unknown or in doubt. Can you identify the author or must the text remain anonymous? Can the date or influencing circumstances of the text be identified?

- Computational response

Count words, frequencies, patterns. Many successes.

- Disciplines: stylometry, stylometrics, author attribution, forensic linguistics. Humanities. *Literary and Linguistic Computing*

Stylometry

From <http://www.soi.city.ac.uk/~peters/research.html>

Stylometric analysis is primarily used to establish authorship of disputed texts. There are perhaps two main categories of stylometric analysis:

- lexical or non-lexical techniques - various ideas have been tried out including word length frequencies, sentence length, letter collocations, but the most reliable seem to be words based methods.
- content against function words - there is strong case for using function words (i.e. those small ubiquitous words that we take for granted). There is some evidence based on speech patterns of aphasia patients that suggests that function words are handled separately by the brain and hence may be less susceptible to variability due to conscious stylistic changes.

See also: <http://www.cs.queensu.ca/achallc97/papers/s004.html>

Generalizing: Texts as Predictors

- Proprietary methods developed by SOK
- Demonstrated: weak (low power) but statistically significant predictive ability of texts, using broadly stylometric methods.
- Opportunity for testing in P^2 context

The Product Placing (P^2) Problem: Computational Response? Sizatola

- Recall: Lesson #2 from PageRank: Exploit work by others
- Classification schemes: a new source of work to be exploited
- Sizatola intuition: Search on the terms describing the properties of C, and find relevant documents. Classify the documents (the hits) using a pertinent classification scheme. Identify promising topics or areas by the density of hits. “Process and System for Matching Products and Markets.” Patent pending, 60/329,703. Inventors: Steven O. Kimbrough, Ian MacMillan, John Ranieri.

The Product Placing (P^2) Problem: Computational Response?

- Example: 1,3-propanediol (PDO), and the USPTO classification scheme
- At random, idea number 66: Personal Care/Cosmetics.

Component in water-absorbent resin composition. Water-absorbent resins absorb body of liquids and are widely used as components of sanitary materials such as disposable diapers, sanitary napkins and incontinence pads. PDO is a suitable surface-cross-linking agent, which enhances the water absorption properties of the composition.

Exemplary Documents

- “Exemplary Documents: A Foundation for Information Retrieval Design,” (David C. Blair & Steven O. Kimbrough), *Information Processing and Management*, vol. 38, 2002, pp. 363–379.
- New and important retrieval concept.
- Evidence favorable but sketchy.
- How to operationalize?

Foundational Operators on Text

$R(\cdot)$ = relevance ranking function

d = a document. t = a (possibly complex) search term

D_s = a group of documents. T_s = a group of search terms

$$(Ds|t)$$

1. Find the documents containing the search term t : $(Ds|t)$

Unranked. Known as boolean retrieval. Easy to implement. Limited value. Easily implemented: Put a K matrix into an RDBMS and exercise SQL.

Also supports proximity searches etc. (with a precursor to K).

$$R(Ds|t)$$

2. Find the documents containing the search term t and rank them by relevance (to the concept associated with t): $R(Ds|t)$

Relevance-ranked document retrieval. Algorithms? DCB. LSI. etc. With DCB, can be read off of the M matrix as described above.

$$R(Ts|t)$$

3. Find the terms associated with term t and rank them by strength of association: $R(Ts|t)$.

Use term-term co-occurrence data. “Empirical Thesaurus” via DCB, using the L matrix. Note: use of thesauruses, faceted indexing as in the AAT (Art and Architecture Thesaurus).

<http://www.getty.edu/research/tools/vocabulary/aat/>

<http://www.britac.ac.uk/portal/h11/aat.html>

$$R(D_s|d)$$

4. $R(D_s|d)$. Rank the documents D_s by relevance to document d .

In terms of the DCB representation this is just $J = K^T \cdot K$.

$$R(Ts|d)$$

5. $R(Ts|d)$.

In terms of the DCB representation this is just

$$I = J \cdot K^T = K^T \cdot K \cdot K^T = M^T.$$

Combinations

- With DCB (and other representations, possibly) stored in an RDBMS, it is straightforward to combine fundamental operators.
- Add: filtering, e.g., find and rank the documents relevant to t including only documents that do not contain the term t' .
- Add, e.g., find documents relevant to search term t that also have terms t_1 and t_2 appearing in the same paragraph.

Conclusion

Much more to say, but we have to stop sometime.

- P²: alive and well; promising results and primary method
- Many promising tools and methods available
- Needed: think through, implement, and test: methodology+support system environment using tools and methods

Reviewing & Previewing

1. concepts – Above; Information Recovery (vs Retrieval); P^2 subproblems (e.g., Topic Extraction: find & describe topics—products, industries, specifications, requirements, signatures—in the data/texts)
2. algorithms – For IR (above and more); visualization
3. data – Data; text collections; USPTO; other sources;
4. knowledge-base – Value added onto public data; proprietary; e.g., for Topic Extraction, implement Exemplary Documents concept (Blair & Kimbrough)
5. experience – With doing studies; experiments and validation exercises
6. tools – Visualization; KWIC/concordance; etc.; workbench

Next Steps

- Jointly design a joint R & D & V(alidation) project to explore, extended, develop and test these ideas and to attack the P^2 problem, including subproblems.
- R & D & V project: clear milestones and deliverables and opt out points; initial phase of about one year; joint work; clear demarcation of rights to use IP.
- Preceded by a joint project design and planning phase, lasting 1–2 months.

Appendix: Sample of Resources (Mostly) on the Web

- Discovery support systems; Gordon+Dumais paper
“Using Latent Semantic Indexing for Literature Based Discovery”
(*JASIS*, 1998, pp. 674–685).
- Thesauruses, faceted indexing (AAT), query expansion.
http://www.boxesandarrows.com/archives/Facets_CV/Bibliography.htm
- Knowledge organization:
http://www.ucl.ac.uk/SLAIS/research/ko_research.htm

- Universal Decimal Classification:

<http://www.udcc.org/>

- eXchangeable Faceted Metadata Language.

<http://xfml.org/>

- Dublin Core

- Advances in Classification Research

http://www.asis.org/Publications/bookstore/sig_cr_1.html

- D-Lib magazine

<http://www.dlib.org/>

- American Society for Information Science and Technology

<http://www.asis.org/>

- North American Industry Classification System (NAICS)

<http://www.census.gov/epcd/www/naics.html>

- UNSPSC: United Nations Standard Products and Services Code

<http://159.169.222.51/eVASPSC/UNSPSC.asp>

From http://www.b2nb.com.au/whatis_UNSPSC.asp:

The UNSPSC stands for the Universal Standard Products and Services Classification (UNSPSC) Code organization. The UNSPSC was created when the United Nations Development Program and Dun & Bradstreet merged their separate commodity classification

codes into a single open system. The UNSPSC Code is the first coding system to classify both products and services for use throughout the global marketplace.

. . . UNSPSC is a classification code (grouping similar products), and is not a product code (describing the product and its specifications).

See also <http://www.industryhub.net/unspsc.htm>

- Product Code Classification Database

This database contains medical device names and associated information developed by the Center for Devices and Radiological Health (CDRH) in support of its mission.

At: <http://www.fda.gov/cdrh/prodcode.html>

- OFFICE OF REGULATORY AFFAIRS (ORA) PRODUCT CODE BUILDER

At: <http://www.accessdata.fda.gov/SCRIPTS/ORA/PCB/PCB.HTM>

- Uniform Code Council

<http://www.uc-council.org/>

- Universal Product Code (UPC) and EAN Article Numbering Code (EAN) Page

<http://www.adams1.com/pub/russadam/upccode.html>

- DIALOG Product Code Finder

<http://library.dialog.com/bluesheets/html/bl0413.html>

- MILSPEC

<http://www.dscc.dla.mil/Programs/MilSpec/>

From:[http://www.dscc.dla.mil/Programs/MilSpec/
ListDocs.asp?BasicDoc=MIL-DTL-13531](http://www.dscc.dla.mil/Programs/MilSpec/ListDocs.asp?BasicDoc=MIL-DTL-13531)

Hose, Rubber and Hose Assembly, Rubber (Hydraulic, Pneumatic, Flexible) FSC 4720

This specification covers the performance requirements and tests for wire-reinforced rubber hydraulic hose and hose assemblies. Hose and hose assemblies covered by this performance specification are intended for use in medium and high pressure hydraulic systems at temperatures between -65 to +200 F. The hose and hose assemblies covered by this specification are military unique because they must be able to operate satisfactorily in temperatures ranging from -65 to +200 F. Commercial products do not operate at these extremes.

Points of contact:

Construction Group Phone: 614-692-1568, 8781 Email:
Construction@dscc.dla.mil

- MILSPEC

<http://dodssp.daps.mil/>

<http://assist2.daps.dla.mil/quicksearch/>

http://stinet.dtic.mil/str/dodiss4_fields.html

<http://www.ihs.com/standards/military-specifications/historical->

- Product Classification Systems

<http://faculty.philau.edu/russowl/product.html>

- Schedule B Export Codes

<http://www.census.gov/foreign-trade/schedules/b/index.html>

About Schedule B Codes... There are millions of trade transactions occurring each year. These transactions are classified under

approximately 8,000 different products leaving the United States. Every item that is exported is assigned a unique 10-digit identification code. Every 10-digit item is part of a series of progressively broader product categories. For example, concentrated frozen apple juice is assigned a 10-digit identifier that is aggregated into a broader category assigned a 6-digit identifier described as apple juice. The 6-digit identifier described as apple juice is aggregated into a broader category assigned a 4-digit identifier described as fruit juices and vegetable juices, etc. The 4-digit identifier is further aggregated into a broader category assigned a 2-digit identifier described as Preparations of Vegetables, Fruit, Nuts etc.

- “product code”, “product classification”
- From: <http://www.computer.org/intelligent/ex2001/x4086abs.htm>

Call for Participants: The E-Commerce Product Classification Challenge

Ellen Schulten, Hans Akkermans, Guy Botquin, Martin Drr, Nicola Guarino, Nelson Lopes, Norman Sadeh This article launches an international research challenge in the area of intelligent e-business. The challenge is to come up with a generic model and working solution that can semiautomatically map a given product description between two different e-commerce product classification standards.

- OFFICE OF MANAGEMENT AND BUDGET

Economic Classification Policy Committee; Initiative to Create a Product Classification System, Phase I: Exploratory Effort to Classify Service Products

AGENCY: Office of Management and Budget, Executive Office of the President

ACTION: Proposed Development of a Comprehensive and Integrated
North American Product Classification System

<http://www.whitehouse.gov/omb/fedreg/napcs1.html>

\$Id: open-tech-foils.tex,v 1.3 2003/05/15 05:44:32 sok Exp \$