

Information from Text: Decision Support for Product Matching*

Steven O. Kimbrough
University of Pennsylvania
565 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, PA 19104-6340
<http://opim-sun.wharton.upenn.edu/~sok/>
kimbrough@wharton.upenn.edu, 215-898-5133

April 20, 2004

*File: dss-product-matching-foils.pdf. See also, May 2003 briefing at <http://opim-sun.wharton.upenn.edu/~sok/asadai/>

Goals of Meeting; Purpose of Foils / Talking Points

- At inception of DuPont–Sizatola project to design and assess a systematic approach to *product matching* problems.
- Project premise: combine expertise and tools from DuPont and Sizatola team. Roughly: DuPont: visualization. Sizatola: information extraction techniques for text.
- Today: discuss concepts and approach; illustrate results and assets; frame problem for follow on discussion and advancement of the project. High-level, non-technical. Also, see previous briefing.
- Product matching: (i) product placing, and (ii) product finding.

Focus, Philosophy, Expectations

- Focus: information from text.
- Philosophy: decision support; humans essential and always 'in the loop'
- Expectations: decision support, not automation; no silver bullet; substantial improvements likely; revolutionary improvements possible, likely requiring aggressive construction of knowledge bases and algorithms. Value of speed as well as power. Also: project organization.

Examples: knowledge bases of information about all known products or about all published product specifications and requirements; extensive knowledge engineering to create a base of heuristic rules for product matching. (NB: all to be assessed)

Central Rôle of Heuristics

- Multiple, related meanings and uses. From the Wikipedia (<http://en.wikipedia.org/wiki/Heuristic>):

The word comes from the same Greek root as “eureka”, meaning “to find”. A heuristic for a given problem is a way of directing your attention fruitfully to a solution. It is different from an algorithm in that it merely serves as a rule of thumb or guideline, as opposed to an invariant procedure. Heuristics may not always achieve the desired outcome, but can be extremely valuable to problem-solving processes. Good heuristics can dramatically reduce the time required to solve a problem by eliminating the need to consider unlikely possibilities or irrelevant states.

- The programme is to identify useful heuristics, and then provide computerized support for applying them.

Example of a Product Matching (Placing) Heuristic

P is our product, the component product we wish to place.

Heuristic 1. [Uses of Similar Component Products] *If P is similar to product Q , and Q is used for X , then P may be useful for X .*

- Operationalization? Many opportunities.
- Example process: identify important attributes of P , retrieve documents matching these attributes, identify the products in those documents, and see what they are used for.
- Issues include: Determining attributes, measuring similarity, finding appropriate document collections, minimizing 'read time' for analysts.

Kinds of Support (in This Context)

- Retrieval
- Extraction
- Thesauruses and query expansion
- Classification
- Heuristic matching
- Visualization

Retrieval

- Standard, boolean retrieval. May include: proximity, phrases, etc.

Useful as far as it goes. Known to be generally poor at finding relevant documents. Plato problem. Does not do ranking. We presently have Python code that is easily modified for any reasonable text query. Plugging on a user interface would allow the code to be mode rapidly exploited by analysts.

- Ranked retrieval. Various methods available. Limited scope of Google, etc.

On project: examine open-source software; DCB (our method; see pages 8–18 of May 2003 briefing); Latent Semantic Analysis; Kolmogorov complexity approach (discuss our work).

Extraction

- Given a set of relevant documents, extract key information from them.
- Value lies in speeding up analysis, offloading humans.
- KWIC, concordances
- Our *specialized concordance*. See example at <http://opim-sun.wharton.upenn.edu/~sok/sizatola/lamineer/kwicreports/>.
- Discuss: generalization and extension of this concept. Link to *information extraction* techniques. Link to POS (part of speech) tagging. Example on next page. See <http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/>

Example POS Tagging with CLAWS

| | | | | |
|---------|-----|---------------|----|---------------|
| 0000194 | 010 | The | 93 | AT |
| 0000194 | 020 | problem | 03 | NN1 |
| 0000194 | 030 | of | 93 | IO |
| 0000194 | 040 | delamination | 06 | NN1 |
| 0000194 | 050 | caused | 98 | VVN |
| 0000194 | 060 | by | 93 | [II/99] RP%/1 |
| 0000194 | 070 | aggressive | 03 | JJ |
| 0000194 | 080 | inks | 93 | NN2 |
| 0000194 | 090 | and | 93 | CC |
| 0000194 | 100 | environmental | 03 | JJ |
| 0000195 | 010 | moisture | 03 | NN1 |
| 0000195 | 020 | is | 93 | VBZ |
| 0000195 | 030 | addressed | 98 | VVN |
| 0000195 | 040 | by | 93 | [II/99] RP%/1 |
| 0000195 | 050 | an | 93 | AT1 |

Thesauruses and Query Expansion

- How to be comprehensive about the search terms? About the products and industries identified?
- Existing thesauruses and dictionaries in electronic form potentially of great value here. Note: WordNet is freely available.
- Can also create an *Empirical Thesaurus* (our term) from collections of relevant documents. We have software for this; see May 2003 briefing: *L* matrix of DCB and Homer, pages 24–8.

Classification

- Useful for focusing on more promising areas.
- Labels in categories may be quite helpful.

Heuristic 2. [Sizatola Example] *Given many documents similar to P in one or more related categories, then likely the uses of the associated products bear specially attending to.*

browser_lamineerU587_s/

(8) class 438: SEMICONDUCTOR DEVICE MANUFACTURING: PROCESS
(3) class 257: ACTIVE SOLID-STATE DEVICES (E.G., TRANSISTORS,
SOLID-STATE DIODES)

Heuristic Matching

- On beyond retrieval. Use of rules, combining matches, and exploiting document structure.
- See NSF proposal, “Automated Regulatory Compliance Support,” by Kimbrough & Lee.

See example rules on next foil, the Kimbrough & Lee proposal.

- Potential benefit lies in speeding the analysis process by focusing.

Example Rules for Heuristic Unification

Expression 1. *If a firm is trading with a proscribed country (P_1) and the firm qualifies as American (P_2), then the firm is in violation of Rule XYZ (P_3).*

Expression 2. *If a firm trades with Iran (Q_1), then the firm trades with a proscribed country (P_1).*

Expression 3. *If a firm has one or more American citizens serving on its board of directors (R_1), or serving as senior executives (R_2), then for the sake of Rule XYZ the firm counts as an American firm (P_2).*

Expression 4. *If a firm is owned more than 30% by American citizens and/or American firms (S_1), then for the sake of Rule XYZ the firm counts as an American firm (P_2).*

Visualization

- Complements categorization. Useful for quickly finding targets of focus.
- Two kinds of query: record-oriented and pattern-oriented.
- Pattern-oriented queries, think: plotting the data. Distribution of one variable? Relationships among two or more variables?
- Our work: Homer idea, Dworman Ph.D. thesis. See May 2003 briefings, pages 24–8. See our JASIS paper. Also, MOTC papers. See: <http://opim-sun.wharton.upenn.edu/~sok/asadai/>.
- DuPont work: Serendipity and Equinox systems.

Just One (or so) Heuristic

- All this and potentially much more in support of just the first heuristic.
- More includes:
 - Integration of capabilities into a smoothly operating system (viz., this project).
 - Construction of supporting knowledge bases, including products, industries, classification schemes, thesauruses, etc.
 - Support for follow-on activities, e.g., market sizing.

Examples of Other Heuristics

Heuristic 3. [Similar Uses] *If P may be useful for X and X is similar to Y, then P may be useful for Y.*

Heuristic 4. [Related Firms] *If P may be useful for X and X is made by firm F and Y is made by firm F, then P may be useful for Y.*

Heuristic 5. [Market Identification] *If P may be useful for X and X is a product in the M market or industry, then P may be useful for other products in M.*

Heuristic 6. [Product Finding] *If Q meets the requirements for specification S, and P is similar to Q, then P may be able to meet the requirements for specification S.*

Heuristic 7. [Competition Matching] *If Q competes with P in some market M , then if Q is useful for X in any market, so P may be useful.*

Comments

- Technologies and techniques relevant for heuristic #1 are also largely relevant for other heuristics.
- Immense, open-ended problem. It is, however, amenable to incremental progress.
- Clearly: good support environments with existing, known technologies will add significant value, if only for speed and organization.
- Add in: specialized knowledge collections, e.g., USPTO patents in HTML, MIL SPEC; results of on-going collection of data, e.g., on products.

\$Id: dss-product-matching-foils.tex,v 1.2 2004/04/20 15:35:59 sok Exp \$