Reactive Chemical Hedges: A Search Tool for Comprehensive Retrieval of Chemical Safety Data[†]

GERALD L. CURNUTT*

Technical Information Services, The Dow Chemical Company, Midland, Michigan 48667

KIRK L. CURNUTT

English Department, Louisiana State University, Baton Rouge, Louisiana 70803

Received July 19, 1990

A step-by-step procedure has been developed for compiling search hedges and has been applied to the topic of reactive chemicals. Special problems encountered in compiling the reactive chemical hedges are presented, including improvements in processing efficiency and recall performance achieved by trimming the hedge with a controlled vocabulary. A reactive chemical value index has been developed and used to rank six chemical databases according to their safety content. From these rankings, new on-line sources have been identified for reactive chemical information.

INTRODUCTION

Since their development by the National Library of Medicine in the early 1970s, hedges have been an obscure method of on-line information retrieval. Hedges are search terms grouped according to concept, stored on a computer, and made accessible to searchers. They allow a searcher to transcend the limitations of the normal saves because hedges provide for variations in spelling and synonyms that the more keyword-oriented saves cannot.[1] Though hedges would seem to possess obvious advantages—not only saving time, and therefore saving money, but also assisting searchers with unfamiliar technologies by providing appropriate search terms—the use of hedges has not fulfilled their potential utility. Hedges have been accused of being insufficiently comprehensive and of promoting "lazy searching"[2]—supposedly because most searchers process the hedges without examining and modifying their contents.

The purpose of this paper is twofold: First, to encourage the continued use of hedges by outlining a step-by-step procedure for their compilation. This procedure will be used to show how the database technology may be applied to increase efficacy of information retrieval, in this case, in the area of reactive chemicals. Second, the procedure will be used to illustrate the potential utility of a new database characteristic, the reactive chemical value index, an index that was specifically developed through the application of the hedge.

PROCEDURE FOR HEDGE COMPILATION

The steps outlined in Figure 1 and described below are adapted from two key papers on the use of hedges in information retrieval. The procedure on hedge compilation was adapted from an article by Dolan,[1] while the hedge-trimming technique described in step C was reported by Sievert and Boyce.[3]

(A) Compiling Search Terms. The first step in compiling a new hedge requires locating an authoritative reference material on the topic of interest. This reference material may be a user's manual or even a dictionary or thesaurus. In the case of the reactive chemical hedge, *Handbook of Reactive Chemical Hazards* by L. Bretherick[4] was selected as the source material. With the aid of this handbook, entire potential search terms were gathered and grouped according to eight categories (chemical reactivity, flammability, dust explosions, shock sensitivity, flash points, autoignition temperatures, vent sizing, and a general category). The number of search terms in each category ranged from 2 to 24, and they included phrases such as runaway reactions, adiabatic calorimetry, and flame temperature. Overall, 64 search terms totaling over 100 words were compiled.

The incorporation of a controlled vocabulary is a vital step in the compilation process. Narrowing the potential search terms as much as possible will save valuable on-line processing time. Any codes or subject headings appropriate for the controlled vocabulary should be included at this step. The controlled vocabularies selected for incorporation should represent all of the files that will be used to process the hedge.

(B) Performance Testing. Unfortunately, new hedges are generally too long and inefficient. The reactive chemical hedge outlined above was first processed in the CA File[5] in May 1988, and the results were less than encouraging. Essentially, the recall was so voluminous that the system limits were exceeded, and the hedge failed to process to completion. In general, the larger the subject (such as reactive chemicals) the more likely the necessity of needing to simplify the initial hedge to ensure that it will process efficiently in the file. During the performance testing, the recall/precision characteristics of the hedge also need to be measured.

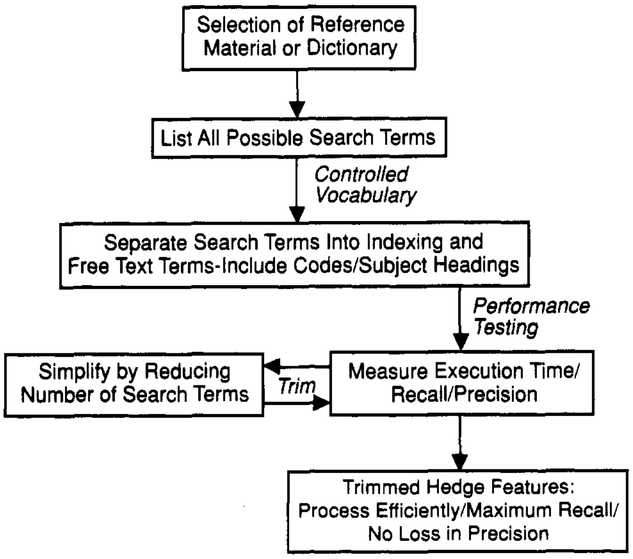
(C) Trimming the Hedge. This simplification procedure, first reported by Sievert and Boyce,[3] has two important aspects: (1) identification of the common components in the search phrases and (2) elimination of the highly posted terms. In the case of the chemical reactivity category, the largest of the eight areas reactive chemicals was divided into, five different phrases referred to the various types of calorimetric measurements that could potentially be used in safety studies. As Figure 2 shows, the common component of these five phrases is the single word calorimetry. Trimming these five phrases to that unbound word will reduce the number of search terms from 13 to one. The other aspect of the trimming procedure of Sievert and Boyce[3] eliminates highly posted terms in search phrases that would require large amounts of computer time to process. As Figure 2 illustrates, flame temperature was one of three search phrases constituting the flammability category. The word temperature happens to be one of the highly posted terms in the CA File. Trimming the three phrases to the single unbound word flame will eliminate temperature as a search term. The elimination of this highly posted term will improve the processing efficiency of the hedge. Figure 3 provides a listing of the search terms constituting the most efficient hedge. As can be seen in Figure 3, the final hedge contains 18 indexing terms and 14 free text terms. The indexing terms represent

[†] Presented at the 22nd Central Regional Meeting of the American Chemical Society, Saginaw, MI, June 6, 1990.

Table I. Retrieval Effectiveness: Precision Studies of Five Reactive Chemical Searches

reactive chemical systems	postings[a]	no. of relevant citations retrieved	precision (%)
3,4-dichloroaniline	41	20	49
polymerization of ethylene oxide	35	24	68
performic acid	21	12	57
polymerization of divinylbenzene	18	10	56
phenol preparation via cumene oxidation	15	9	60

[a] CA File/reactive chemical hedge no. 4.


Selection of Reference Material or Dictionary

List All Possible Search Terms

Controlled Vocabulary

Separate Search Terms Into Indexing and Free Text Terms-Include Codes/Subject Headings

Performance Testing

Simplify by Reducing Number of Search Terms

Trim

Measure Execution Time/ Recall/Precision

Trimmed Hedge Features: Process Efficiently/Maximum Recall/ No Loss in Precision

Figure 1. Procedure for compilation of hedges.

Chemical reactivity category:
- Accelerating rate ***calorimetry***
- Safety ***calorimetry***
- Adiabatic ***calorimetry***
- Differential scanning ***calorimetry***
- Heat flow ***calorimetry***

Trimmed: ***calorimetr?***

Flammability category:
- ***Flame*** propagation
- ***Flame*** temperature
- Cool ***flames***

Trimmed: ***flame?***

Figure 2. Examples of trimming initial reactive chemical hedge.

the controlled vocabulary that was incorporated from the Chemical Abstracts Index Guide.[6] To improve the precision, we also included four negation terms. These terms have been added to exclude information on health hazards, industrial hygiene, pollution, and toxicology.

PERFORMANCE DATA ON THE REACTIVE CHEMICAL HEDGES

Figure 4 illustrates some of the performance data gathered during the development of the reactive chemical hedges. Overall, the hedges were trimmed three times. As Figure 4 illustrates, the first hedge consisting of 64 terms required nearly 14 min to process and retrieved roughly a half-million citations from the CA File. When this initial hedge was reduced to 43 terms, the execution time decreased by more than half whereas the number of postings increased by some 200 000 citations. Hedges 3 and 4, the products of further trimming, continued to show reduced processing time while eventually increasing recall by more than 50% from the first hedge. The more than 800 000 citations retrieved by the fourth hedge in just under 3 min represent roughly 10% of the total CA File.

Database: CA File
Controlled Vocabulary: Chemical Abstracts

Indexing Terms

1. Accident?
2. Calorimetr?
3. Deflagration
4. Detonat?
5. Explosi?
6. Fire?
7. Flame?
8. Flammab?
9. Flash Point?
10. Hazard?
11. Heat?
12. Ignition
13. Injur?
14. Reactiv?
15. Rupture
16. Safe?
17. Thermal Analysis
18. Thermodynamics

Free Text Terms

19. AIT
20. Dangerous
21. Drop Weight
22. DSC
23. DTA
24. Exotherm?
25. Impact Sensitivit?
26. Incident?
27. Runaway?
28. Self Accelerati?
29. Shock Sensitivit?
30. Stabilit?
31. Vent Sizing
32. VSP

Negation Terms

33. Health
34. Hygiene
35. Pollution
36. Toxic?

Figure 3. Reactive chemical hedge no. 4 search term listing. The controlled vocabulary was from ref 6.

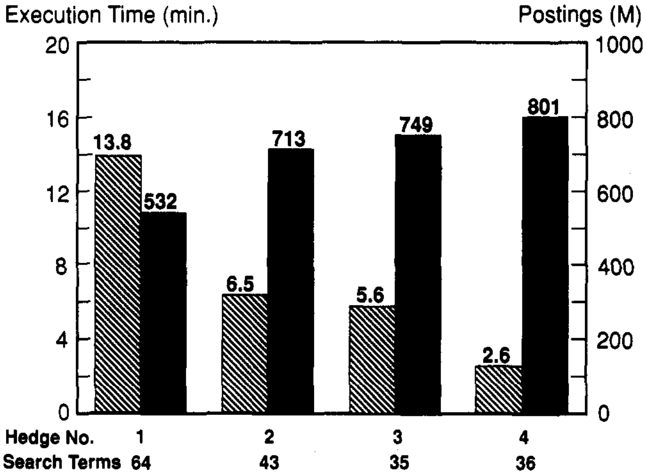


Figure 4. Processing efficiency and recall for reactive chemical hedges. The execution time is for the CA File; evaluation date 2-5-90. Posting units are thousands of citations.

Precision is as equally an important performance parameter as processing efficiency and recall. Sievert and Boyce[3] have indicated that precision could be improved by simplifying the hedge with their trimming technique—a surprising conclusion since trimming tends to maximize recall. The results obtained in this study indicate that the precision attained with the reactive chemical hedge is only moderate. As outlined in Table I, reactive chemical searches were performed on five chemical systems by combining the search terms of the fourth hedge with the appropriate chemical synonyms, CAS Registry Numbers, and descriptors for the specific processes, i.e., po-

```
=>FIL REG COST = 1831

=>S PHENOL/CN
L1              1 PHENOL/CN

=>S CUMENE/ CN
L2              1 CUMENE/CN

=>S CUMENE HYDROPEROXIDE/CN
L3              1 CUMENE HYDROPEROXIDE/CN

=>FIL CA

FILE 'CA' ENTERED AT 11:38:39 ON 13 SEP 90

=>S  L1/P AND (L2(L) OXIDN OR L3)
L4              166 L1/P AND (L2(L) OXIDN OR L3)

=>S L3 AND (L2(L) OXIDN)
L5              200 L3 AND  (L2(L) OXIDN)

=>ACT REACHEM /Q
L6             QUE (EXPLOSI? OR DETONAT? OR ACCIDENT? OR DEFLAGRATION OR
               INJUR? OR RUPTURE OR HAZARD? OR SAFE ? OR THERMODYNAMICS
               OR THERMAL ANALYSIS OR REACTIV? OR CALORIMETR? OR HEAT?
               OR FLASH POINT? OR FIRE? OR FLAMMAB? OR FLAME? OR IGNITION
               NOT (TOXIC? OR HEALTH OR HYGIENE OR POLLUTION))

=>ACT REACHEM1/Q
L7             QUE (DANGEROUS OR STABILIT? OR INCIDENT? OR RUNAWAY? OR
               SELF ACCELERATI? OR EXOTHERM? OR DSC OR DTA OR AIT OR VSP
               OR VENT SIZING OR SHOCK SENSITIVIT? OR IMPACT SENSITIVIT? OR
               DROP WEIGHT NOT (TOXIC? OR HEALTH OR HYGIENE OR POLLUTION))
               /AB,BI

=>S (L6 OR L7) AND (L4 OR L5)
L8              15 (L6 OR L7) AND  (L4 OR L5)
```

Figure 5. Search strategy for retrieving reactive chemical information on the cumene phenol process from the CA File.

Table II. Reactive Chemical Value Index: Rankings of Six On-Line Chemical Databases

files	postings (M)[a]	file size ($\bar{M}$)	av. no. of chemicals indexed per citation	time span (yr)	reactive chemical value index
CA File	801	9.0	5.5	23	11.3
Chemical Safety Newsbase	5	0.023	1.1	9	2.2
NTIS	301	1.4	0.4	26	2.2
Inspec	438	3.3	0.6	21	1.7
Compendex Plus	459	2.6	0.4	20	1.4
DOE Energy	589	2.4	0.2	16	0.8

[a] Reactive chemical hedge no. 4: postings on 2-7-90.

lymerization or oxidation. The search strategy used to retrieve citations on the reactivity hazards associated with the preparation of phenol from the oxidation of cumene is listed in Figure 5. From the data presented in Table I, the percent of relevant citations as compared to the total number of retrieved citations averaged 60. A precision of this magnitude seems inevitable considering the exceedingly high recall attained with this hedge in the CA File.

THE REACTIVE CHEMICAL VALUE INDEX

To this date, the selection of a suitable database for reactive chemical searches has been problematic, with criteria generally as subjective as experience and simple database descriptions. We wish to report here a new database characteristic, the reactive chemical value index, that should assist searchers in locating relevant safety databases. The value index developed is an empirical parameter that constitutes a scientific formula devised to evaluate the chemical safety content of any particular database.

The calculation of the reactive chemical value index (RCVI) is relatively simple. The equation is as follows:

$$\mathrm{RCVI} = \left(\frac{\begin{array}{c}\text{postings}\\\text{(reactive chemical hedge)}\end{array}}{\text{file size}}\right)\left(\begin{array}{c}\text{av no. of}\\\text{chemical substances}\\\text{indexed per citation}\end{array}\right)\left(\begin{array}{c}\text{time span}\\\text{in years}\end{array}\right) \quad (1)$$

Dividing the postings of the hedge by the file size provides a measure of the probability of finding a reactive chemical citation in the database. This probability factor is then multiplied by the average number of chemical substances indexed per citation. Values were determined for this measure of chemical substance content from information provided by the database producers and are shown in Table II. These numbers are then multiplied by the time span in years that the database covers.

The reactive chemical hedge was processed in several databases drawn from a study of Cipra and Damron[7] that identified 50 on-line databases potentially useful for retrieving safety information. Table II shows the rankings of six chemical databases that were selected for further study based on the number of postings to the hedge. The CA File has the highest ranking with a value index equal to 11.3, which is about 5 times larger than that for any of the other five databases. The surprising result provided by the rankings is that NTIS,[8] Inspec,[8] and Compendex Plus[8] all have potential nearly equivalent to the Chemical Safety Newsbase file. These three databases have previously not been considered valuable sources of reactive chemical information.

Figure 6 shows the results from five searches that support the index rankings provided in Table II. The bar chart in Figure 6 shows a plot of the number of relevant citations retrieved when the five reactive chemical searches outlined in

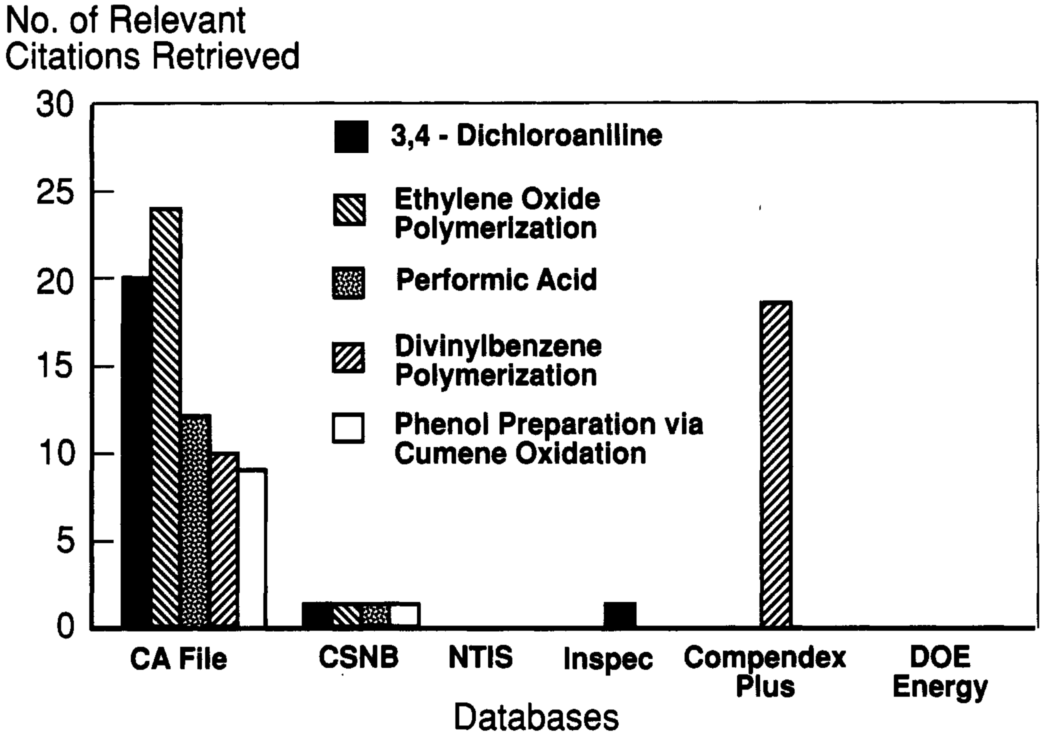


Figure 6. Multidatabase retrieval for five reactive chemical searches.

Table I and Figure 5 were processed in each of the six databases. As seen in Figure 6, the CA File provided the most useful information. We should also point out that none of the citations retrieved from the CA File and Chemical Safety Newsbase overlapped. The excellent potential of these two sources is reflected in the index rankings provided in Table II. The usefulness of Inspec and Compendex Plus as reactive chemical sources is indicated by the number of citations retrieved for two of the five searches. For example, Inspec provided an additional useful citation on 3,4-dichloroaniline. Several citations were retrieved from Compendex Plus on the polymerization hazards of divinylbenzene, very few of which overlapped with the citations found in the CA File. While the values of the index do not show a consistent correlation with the number of relevant citations retrieved, they do indicate a degree of potential usefulness of these databases, as indicated by the search results.

CONCLUSION

With the increasing diversity of information sources, the need to use hedges to retrieve relevant citations and provide an empirical parameter to help evaluate those sources seems evident. The step-by-step procedure for compiling hedges outlined in this paper will provide searchers with a tool for enhancing their productivity and avoiding the accumulation of costly on-line charges. Eecause hedges are a relatively unexplored area of information science, the continued study and further development of their potential will only increase the utility of this search tool. In addition, the hedge may prove valuable in analyzing and quantifying the contents of databases—such as in the reactive chemical value index whose development has been outlined here. Therefore, searchers should work closely with the database producers and vendors to further develop this method of on-line information retrieval.

ACKNOWLEDGMENT

We thank Marcia Dilling for her valuable comments and criticisms. We gratefully acknowledge the support of this work by The Dow Chemical Company.

REFERENCES AND NOTES

(1) Dolan, Donna R. Hedges for Online Searching. *Database* **1980**, *3* (1), 79–82.

(2) Tilley, Carolyn. National Library of Medicine. Private Communication, Feb 1990.

(3) Sievert, MaryEllen; Boyce, Bert R. Hedge Trimming and the Resurrection of the Controlled Vocabulary in Online Searching. *Online Rev.* **1983**, *7* (6), 489–494.

(4) Bretherick, L. Handbook of Reactive Chemical Hazards, 3rd ed.; Butterworths: London, 1985.

(5) CA File is a service mark of Chemical Abstracts Service.

(6) *Chemical Abstracts Index Guide*; American Chemical Society; Chemical Abstracts Service: Columbus, 1987.

(7) Cipra, David M.; Damron, C. Frazier. Safety, A Guide to Nearly 50 Databases Containing Occupational, Personal and Other Safety-Related Information. *Database* **1985**, *8* (2), 23–30.

(8) NTIS is a registered trademark of National Technical Information Service; Inspec is a registered trademark of Institution of Electrical Engineers; and Compendex is a registered trademark of Engineering Information, Inc.