Information Processing & Management Vol. 22, No. 6, pp. 487–492, 1986
Printed in Great Britain.

0306-4573/86 $3.00 + .00

A THESAURUS FOR END-USER INDEXING AND RETRIEVAL

GARY W. STRONG† and M. CARL DROTT
College of Information Studies, Drexel University, Philadelphia, Pennsylvania

(*Received* 2 *May* 1986; *in revised form* 8 *July* 1986)

Abstract—Direct end-user data entry and retrieval is a major factor in achieving an economical information retrieval system. To be effective, such a system would have to provide a thesaurus structure which leads novice end-users to browse subject areas before retrieval and yet provides control and coverage of terms in a domain. A faceted hierarchical thesaurus organization has been designed to accomplish this goal.

THE PROBLEM

One of the biggest problems in engineering research and development (R&D) in a highly competitive area is the ability to assemble pertinent literature, proprietary company information and competitor's product information in a timely fashion. Yet a traditional information retrieval system requires specialist personnel and a time interval between when information is selected for entry into the system and the time a user's queries may be satisfied. End-users cannot easily retrieve information about desired topics without sufficient search experience, and information retrieval databases are not easily updated by end-users.

In a traditional information retrieval system, professional abstracters and indexers, using terms from a controlled-vocabulary thesaurus, generate terms by which records are indexed in order to jointly maximize precision (the fraction of the number of documents retrieved that are relevant) and recall (the fraction of the total number of relevant documents in the collection that are retrieved)[1]. During searches, the thesaurus is used to generate search terms. A search strategy is formulated by an information science specialist, using Boolean search techniques (conjunctions, disjunctions and textual adjacencies among terms). The number of items found provides feedback to refine the search. The cost of such an approach in terms of both salary and time of professional staff, such as abstracters, indexers and information science specialists, makes it impractical in many field applications. This cost factor becomes critical in a competetive industrial R&D setting, where there is an emphasis upon rapid development.

Some information retrieval applications involve an equally impractical full-text searching methodology. The complete text of each record is fully indexed, except for a list of "stop-words" which are useless in performing searches. Even though the need for professional staff is less than in the case outlined above, searches performed on such a natural language, full-text index typically demonstrate poor precision and recall[2]. Although the end-user can learn techniques to iteratively improve precision and recall, the results are highly dependent upon the end-user's skill in recalling related words and their relationships in the desired subject area as well as upon the content behavior of the subject area, such as its boundedness or hierarchical structure.

A low-cost, practical information retrieval system, if it were to be designed, would require a thesaurus, but one in which end-users would be able to browse research topics by means of an organization which is concept-based rather than term-based as is the typical thesaurus. Such an information retrieval system would also allow end-users, indeed encourage them, to enter information directly into the database. These features would negate the need for costly abstracting and indexing staff. In addition, the retrieval language would have to support direct user interaction. There are few approaches today, however,

†To whom correspondence should be addressed.

Table 2—Effect of Cyclic AMP on the Incorporation of [1-^{14}C]Acetate & [2-^{14}C]Malonate into Lipids by Chicken Liver

[Experimental conditions are as described in Table 1, except that the concentration of malonate or acetate used was 2.5 m*M*. Values are average ± SD of six experiments]

Substrate	nmoles incorporated into lipids/100 mg tissue/hr		Inhibition %
	−cAMP	+cAMP	
[1-^{14}C]Acetate	143.6 ± 34.5	58.63 ± 22.9	59 ± 4.6
[2-^{14}C]Malonate	60.6 ± 17.2	28.8 ± 1.7	53 ± 4.8

things, its activity has been shown to be regulated by allosteric effectors and also through covalent modification involving phosphorylation-dephosphorylation reactions. In recent years, contradicting reports appeared regarding the role of acetyl-CoA carboxylase in glucagon and cyclic AMP effected inhibition of fatty acid synthesis. Thus injection of glucagon[10] or its addition to isolated liver cells[11] was shown to result in a drastic reduction in the activity of acetyl-CoA carboxylase. Further, addition of dibutyryl cyclic AMP to liver slices or to cell-free systems also lead to a significant decrease both in the enzyme activity and fatty acid synthesis. In contrast, however, Cook *et al.*[12] reported recently that while the injection of glucagon to rats resulted in 60% inhibition of fatty acid synthesis in 15 min, no significant change in the activity of acetyl-CoA carboxylase was observed. It is also worthwhile to note that Carlson and Kim[13,14] failed to obtain any effect of cyclic AMP on ATP-dependent inactivation of purified rat liver acetyl-CoA carboxylase. In the light of these controversial reports, the results presented here are of considerable significance since they show that the site of cyclic AMP action is unlikely to be acetyl-CoA carboxylase and suggest that it is probably distal to the location of this enzyme in the fatty acid synthetic pathway from acetate.

This work was supported by University Grants Commission, New Delhi.

References

1. Rous, S. (1970) *FEBS Lett.* **12**, 45-48.
2. Allred, J.B. & Roehrig, K.L. (1973) *J. biol. Chem.* **248**, 4131-4133.
3. Akhtar, M. & Bloxham, D.P. (1970) *Biochem. J.* **120**, 11p.
4. Bricker, L.A. & Levey, G.S. (1972) *J. biol. Chem.* **247**, 4914-4915.
5. Bhat, N.R., Madhava Rao, A. & Murthy, S.K. (1978) *Indian J. Biochem. Biophys.* **15**, 39-42.
6. Goodridge, A.G. (1973) *J. biol. Chem.* **248**, 1924-1931.
7. Capuzzi, D.M., Rothman, V. & Margolis, S. (1974) *J. biol. Chem.* **249**, 1286-1294.
8. Harris, R.A. (1975) *Archs Biochem. Biophys.* **169**, 168-180.
9. Lane, M.D., Moss, J. & Polakis, S.E. (1974) in *Current Topics in Cellular Regulation*, Vol. 8 (B.L. Horecker & E.R. Standtman, eds.), 139-187, Academic Press.
10. Klain, G.J. & Weiser, P.C. (1973) *Biochem. biophys. Res. Commun.* **55**, 76-83.
11. Muller, P., Singh, A., Orci, L. & Jeanrenaud, B. (1976) *Biochim. biophys. Acta* **428**, 480-494.
12. Cook, G.A., Nielsen, R.C., Hawkins, R.A., Mebiman, M.A., Lakshmanan, M.R. & Veech, R.L. (1977) *J. biol. chem.* **252**, 4421-4424.
13. Carlson, C.A. & Kim, K.H. (1973) *J. biol. Chem.* **248**, 378-380.
14. Carlson, C.A. & Kim, K.H. (1974) *Archs Biochem. Biophys.* **164**, 478-489.

Arginase in Human Saliva

R. GOPALAKRISHNA & B. NAGARAJAN
Department of Microbiology, Cancer Institute, Madras 600020

Manuscript received 24 *June* 1978; *revised manuscript received* 11 *September* 1978

Arginase activity in human saliva was in the range of 180-500 IU/litre, a value which is 400-fold of that found in serum, thereby suggesting a physiological function for arginase. The enzyme was heat-stable, but unstable during cold storage at −10°C or 4°C for a week. At *p*H 9.5 K_m of arginase fgr arginine was 5m*M*. Ornithine showed a mixed type of inhibition. The molecular weight was estimated as 120,000 by gel chromatography.

Although arginase (L-arginine amidinohydrolase, EC 3.5. 3.1) is present in various mammalian tissues like liver, salivary gland, small intestine, kidney and pancreas[1,2], no information appears to be available on the presence and characterisation of arginase in human saliva. The present study is an attempt to remove this lacunae. Arginine-free base, Tris(hydroxymethyl amino) methane, L-ornithine hydrochloride, 2,3-butanedione monoxime and thiosemicarbazide were obtained from Sigma Chemical Co. and ninhydrin from BDH.

An aliquot (3 ml) of saliva was centrifuged at 2,500 × *g* for 10 min. To the supernatant, an equal volume of 0.02*M* Tris-HCl buffer (*p*H 7.5) containing 0.025*M* $MnCl_2$ was added and activated by heating at 55°C for 10 min. This is referred as 'heat treatment' in the text. In all the experiments, this buffer was used unless otherwise stated.

Arginase activity was assayed by incubating 0.1 ml of preactivated saliva with 40 μmoles of glycine-NaOH

RP9

buffer (*p*H 9.5) and 40 μmoles of L-arginine (*p*H 9.5) to a final volume of 1.0 ml at 37°C for 15 min. The reaction was stopped by adding 1.0 ml 10% TCA. One ml of the supernatant was used for urea estimation[3]. Suitable reagent and enzyme blanks were set up. Stoichiometric nature of this reaction was established with $NaHCO_3$-Na_2CO_3 buffer, since glycine interfered with colour product, when ornithine was estimated by Chinard's method[4]. In other experiments, urea was estimated to measure arginase activity. In cases where mercaptoethanol was used, which inhibited urea diacetyl monoxime colour reaction, the estimation of ornithine became necessary. It was also necessary to run standards in the presence of varying $MnCl_2$ concentrations to account for the interference with the colour reaction. Serum arginase was also assayed by the above procedure after subjecting 0.1 ml of serum to centrifuge dialysis using Sephadex G-50 pre-equilibrated with Tris buffer[5]. Agrinase activity in saliva and serum was expressed in IU/litre, where one IU was defined as the amount of enzyme producing one μmole urea/min. In saliva, the presence of arginine desimidase (EC 3.5.3.6) and ornithine carbamyltransferase (EC 2.1.3.3) was checked[6,7].

Molecular weight was determined by gel chromatography using Sephadex G-100 according to Andrews[8]. Two ml of preactivated saliva was applied to 2.5 x 50 cm column and the enzyme eluted with 0.01 *M* Tris buffer (*p*H 7.4) containing 0.1 *M* KCl. For calibration, ribonuclease T_1, myoglobin, chymotrypsinogen, bovine serum albumin, both monomer and dimer, and rat liver arginase were used.

The enzyme preparation produced an equimolar amount of ornithine and urea from arginine confirming total arginase activity. This preparation was free from arginine desimidase and ornithine carbamyltransferase activities. Heat treatment of enzyme in the presence of $MnCl_2$ at 55°C showed a 10% increase over the activity of the native enzyme. The enzyme was stable at 60°C at least for an hour. But in the absence of Mn^{2+}, the enzyme lost about 80% of activity within 15 min at 60°C. Though the presence of Mn^{2+} was necessary during incubation, higher concentrations inhibited enzyme activity as shown in Fig. 1.

The implications of the presence of Mn^{2+} are again brought out by dialysis experiments. The enzyme lost only 20% of its original activity when dialysed in the presence of Mn^{2+} and further heat-treatment did not make any difference. However, about 95% of the activity was lost in dialysis against buffer free from Mn^{2+}. Only 20% of the activity was restored when assayed in the presence of Mn^{2+} and showed 80% of activity when heat treated again with manganese before enzymatic reaction (Table 1).

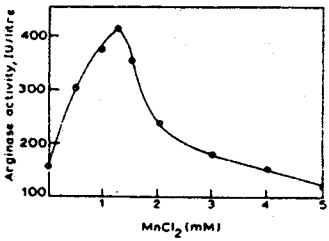

Fig. 1—Effect of varying $MnCl_2$ concentration on arginase activity in saliva [Centrifuged saliva was assayed without further preactivation]

Table 1—Effect of Dialysis on Arginase Activity

[Heat-treated enzyme with initial activity of 425 IU/litre was used for dialysis for 24 hr. Values represent the activity in IU/litre and those in parentheses % initial activity]

Exptl condition	Activity	
	Before heat-treatment	After heat-treatment
Dialysis *vs* Tris buffer (Mn^{2+} present)	335 (78)	341 (80)
Dialysis *vs* buffer (Mn^{2+} absent)		
Assay without Mn^{2+}	18 (42)	—
Assay with Mn^{2+}	89 (21)	328 (78)

Saliva stored at 4°C in the presence of Mn^{2+} lost appreciable activity during the course of a week. Similarly, by freezing and thawing, the enzyme lost 90% of its original activity even in the presence of Mn^{2+} and this loss could not be restored by activation with heat treatment.

The optimum *p*H was 9.7-10.0 and the K_m value for arginine was 5m*M*. There was no excess substrate inhibition or activation. Mercaptoethanol at 5m*M* level had no effect on the reaction. Ornithine showed a mixed type of inhibition (Fig. 2). Most of the enzyme activity was eluted in a single peak from Sephadex G-100 column corresponding to a mol. wt 120,000. These properties were similar to those of several mammalian arginases[9,10].

The enzyme activity of unstimulated saliva in 20 normal adults was in the range of 180-500 IU/litre (The mean value being 370 ± 50). there was no sex variation. The level of salivary enzyme activity was much higher compared to that of human serum arginase (0.5-2.3 IU/litre). In saliva, ornithine carbamyltransferase was not detectable, though it was present in serum. It

thesauri (i.e. sets of controlled terms which show subsets that are related in some lexical way[6]) more effectively in the design of an end-user information retrieval system, during both data entry and data retrieval. Our approach is to conceive of facet relationships in a spatial manner. The dimensions of this "space" are general facets (or "prime dimensions of knowledge," cf. [7]) by which documents may be classified.

USE OF THE THESAURUS

As a result of our design which situates facets at the root of the hierarchy, the user, in classifying a document, is encouraged to choose a number of different hierarchical paths beginning with different broad facets and proceeding to more detailed subfacets. One hierarchical path alone is usually insufficient to fully categorize a document. Even though a searcher need specify only one of the indexed facets to retrieve something, a more effective search should specify a number of facets whose intersection more clearly specifies the target concept. Thus precision is under user control in a way which relates quite clearly to content. The relation is a spatial relation which we claim is easily understood by users because of the transfer of learning from knowledge of three-dimensional physical space. In other words, the more dimensions of the space that one can constrain (and constrain more specifically), then the more precisely one specifies content.

We have found that by relating the thesaurus structure to the design of the human–computer interface, the system is made easy enough to use that end-users can be relied upon to classify their own documents. The thesaurus structure is presented to the user as subfacets under a specific facet or subfacet previously chosen, as one might walk through a tree-structured menu system. By picking facets, the user walks down a hierarchical tree from broad to narrow facets, such as from METHODOLOGY to ANALYSIS to FUNCTIONAL. Even at the narrowest level, facets should represent different aspects of a subject which are not exclusive categories unto themselves. For example, some methodologies may be a combination of both a functional analysis and data analysis but a single article need not address both aspects. This subdivision scheme tries to preserve the idea of division into independent subfacets for as long as possible, however, not at the price of fighting human nature. As we become more specific we tend to need exclusive categories. Therefore, the narrowest subfacets tend to represent this more restrictive kind of classification and are more like terms than facets. For any walk down the hierarchical structure from the root, a facet sequence (such as METHODOLOGY.ANALYSIS.FUNCTIONAL) is created which may be used either during classification or during retrieval.

In order to make precision more directly under user control (or program control in an automated system), each document, upon entry, must be classified under more than one facet of the thesaurus. It should, in fact, be classified under *all* facets deemed appropriate by the enterer. This also insures that someone interested in a particular document can retrieve it by specifying only one of its indexed facets. For example, if a journal article on software engineering has been classified under the following facet sequences:

METHODOLOGY.ANALYSIS.FUNCTIONAL;
TOOLS.ACTION DIAGRAMS;

then someone could retrieve it by specifying either of these facets. Upposting to the broader facets in each of its classification sequences, as is done with terms in traditional information retrieval system, helps to increase recall.

The most important aspect of the thesaurus which we have designed, however, is not that it offers strict control over recall and precision as suggested, but that it not be a barrier to end-user indexing and retrieval of information. In fact, the semiautomated information retrieval system which we are currently developing automatically adjusts intersections of facet sequences until satisfactory retrieval sets are achieved, rather than put control of precision in the hands of the human user. Our motivation and method for doing this will be discussed in a future article. Furthermore, we are also exploring issues relating to the geometry of concepts in the multidimensional concept space, such as different amounts of constraint in different dimensions and their relation to search strategy as spatial schemata.

Based upon our experience thus far in implementing such a thesaurus, however, we have developed some general rules for using a multidimensional, faceted thesaurus for document classification. They are as follows:

(1) Every document must have at least one facet classification. Most documents should have at least two or three.
(2) In assigning thesaurus facets one need not go to the full depth of the thesaurus (that is, one can stop building a facet sequence even though there still remain more specific levels within the thesaurus from which to select subfacets).
(3) A facet sequence can always have an additional level of specificity; but as a rule, users should not be encouraged to do this since the extra time and effort have a very low payoff. On the other hand, because the retrieval system automatically upposts specific facets with the broader facets, it will not hurt retrieval.
(4) The assignment of a facet sequence to a paper neither requires nor precludes the assignment of any other facet sequence to the same paper.

MODIFYING THE THESAURUS

Any thesaurus represents a series of compromises made in order to enhance communication. More specifically, individuals give up their freedom to use exactly the "right" word so that all users can share a common vocabulary. The need for sharing limits precision but enhances the integration of ideas. As a general rule, changes to the thesaurus should represent well-considered group decisions within a specific domain. In our proposed thesaurus, the higher the level of the facet being added or changed, the greater the need for agreement across the research domain. Changes to the thesaurus are also best discussed in terms of positive instances, i.e. specific sample documents to which the new facet should apply. Again, the higher the level being changed, the more samples that should be considered.

In our experience, one of the easiest errors to make in working with such a multidimensional faceted thesaurus is to forget that documents are to be classified with a number of thesaurus facet sequences. No single facet sequence is definitive. It should also be recognized that the major function of the thesaurus is to group documents, not to separate them. Users who search using thesaurus facets will be able to browse quickly the information retrieved and eliminate what is not wanted, but a user cannot even speculate on items that are not retrieved. The hierarchical organization is the foremost tool for this grouping action. Changes to the hierarchy should always be considered in terms of what will be brought together. With this in mind, we have developed general rules for thesaurus modification as shown below:

(1) The addition (or deletion) of facets should arise from the presence (or absence) of actual document samples which "need" to be described by that facet.
(2) In order to follow the concept of independent facets, the same subfacet should not be repeated under different facets.
(3) At any point in the thesaurus, there should be no more than eight to ten choices of subfacets at that level under the given broader facet.
(4) Scope notes which explain the facets help to promote consistency and should be added when possible.
(5) The thesaurus is a research guide as well as an information retrieval tool. The organization and terminology should therefore reflect company standard terminology and practices.

GUIDELINES FOR THE DATABASE ADMINISTRATION FUNCTION

In our view, a thesaurus is a living document which describes the relationship between controlled key terms by which data are classified and by which data may be retrieved. It evolves as the field changes and as users gain experience in a field. It is the responsibility of a database administrator (DBA) to collect users' suggested additions and deletions to

the thesaurus. There should be a mechanism, therefore, for users to suggest their own facets when they do not find the thesaurus facets to be adequate for classifying documents. Additions may, for example, come in the form of error messages which occur when new user documents indexed by facets not in the thesaurus are actually added to the database by the DBA. (The DBA oversees the addition of user records to the database.) Such error messages may be generated by the database software when it does not find a match between indexing facets and existing thesaurus facets. Suggestions for deletions as well should be available from users' reactions to the suitability of existing facets. There should be a mechanism by which the DBA collects such reactions. Which facets should be added or deleted is, however, a judgment decision to be made by the DBA, using the modification rules stated above and under advice from management.

CONCLUSION

A thesaurus for end-user indexing and retrieval has been designed which has an analogy to multidimensional scaling, where each dimension corresponds to a broad facet. Users classify their own documents by locating them in this multidimensional conceptual space, using a hierarchical organization of subfacets on each dimension as a guide. Recall is improved by needing to specify a hierarchical facet sequence along only one of the coded facet dimensions in order to retrieve the desired material. Emphasis is upon classifying documents under multiple dimensions in this "intellectual space," however, because precision is under user control by intersecting facets of interest during search. Guidelines for building and maintaining such a thesaurus are described. Research concerning the foundations for such a multidimensional faceted thesaurus is underway[5] and will be reported at a later date. The purpose of this paper is to discuss a new methodology for thesaurus design which may have implications for efficient information retrieval by novices.

REFERENCES

1. Salton, G. *Automatic Information Organization and Retrieval.* New York: McGraw-Hill; 1968; 283–284.
2. Bernstein, L. M.; Williamson, R. E. "Testing of a Natural Language Retrieval System for a Full Text Knowledge Base." *Journal of the American Society for Information Science* 35:235–247; 1984.
3. Pattee, H. H. *Hierarchy Theory.* New York: Braziller; 1973; 75.
4. White, H. D.; Griffith, B. C. "Author Cocitation: A Literature Measure of Intellectual Structure." *Journal of the American Society for Information Science* 32:163–171; 1981.
5. Strong, G. W.; Whitehead, B. A. "Affordance-Based Inference in Natural and Artificial Systems." Article submitted for publication; 1986.
6. Wang, Y. C.; Vandendorpe, J.; Evens, M. "Relational Thesauri in Information Retrieval." *Journal of the American Society for Information Science* 36:15–27; 1985.
7. Pountain, D. "New Database Ideas." *Byte* 10:389–397; 1985.