Serdica
Bulgariacae mathematicae publicationes

Сердика
Българско математическо списание

A TECHNIQUE FOR COMPARING TERM CLASSIFICATIONS*

P. BOLLMAN, E. KONRAD, H. ZUSE

The purpose of this paper is to describe a software tool which can be used to compare association structures of term classifications, thesauri, and dictionaries. The conceptual work is supported by experiments with the FAKYR document retrieval system which has been implemented on an IBM 370/158.

Introduction. The problem of comparing term classifications has been tackled from different points of view. Some authors have compared classifications intellectually [2, 8] and the evaluation has been conducted by running recall-precision experiments [2, 7, 8]. To what extent different classifications lead to different retrieval results has not yet been answered.

A step towards the solution of this problem is to set up a measure for comparing classifications. In large systems it takes a lot of time to improve classifications intellectually. A tool for detecting deficiencies would be very helpful.

In the area of taxonomy there are approaches dealing with the quantitative comparison of classifications. Hartigan uses distance measures between dendrograms [6], while Anderberg [1] defines similarity measures between partitions. For our purposes these methods have two disadvantages. First, each classification has to be defined on the same set of objects — this is normally not fulfilled for two different thesauri. Second, local deviations cannot be detected automatically. We propose a similarity measure that avoids these two disadvantages. Furthermore it can be used for comparing association networks. A modification of this measure has been applied to bilingual association networks (4).

The similarity measure. Let K_1 and K_2 be two classifications of the set of terms T_1, T_2 resp. If x is a term of $T_1 \cup T_2$, let $N_i(x)$ be the set of terms that are members of the same class of terms (with respect to the classification K_i). $N_i(x)$ can be considered as the set of neighbours of x, each term is neighbour of itself. We now define a local similarity $\alpha_x(K_1, K_2)$ (with respect to x):

$$\alpha_x(K_1, K_2) = |N_1(x) \cap N_2(x)| / |N_1(x) \cup N_2(x)|.$$

We have a good reason to apply the Tanimoto measure between sets [5]. Using the overlap measure $|N_1(x) \cap N_2(x)|$ terms in big classes seem to be more appropriately classified than others.

*Delivered at the Conference on Systems for Information Servicing of Professionally Linked Computer Users, May 232-9, 1977, Varna.

SERDICA Bulgaricae mathematicae publicationes. Vol. 10, 1984, p. 124—127.

Example. $T_1=\{a, b, c, d, e, f, g\}$, $T_2=\{b, c, d, e, f, g, h\}$,
$K_1=\{\{a, b, c\}, \{c, d, e\}, \{d, f, g\}\}$, $K_2=\{\{b, h\}, \{c, d, e\}, \{f, g\}\}$
$N_1(a)=\{a, b, c\}$, $N_2(a)=\emptyset$, $\alpha_a(K_1, K_2)=0$, $N_1(d)=\{c, d, e, f, g\}$, $N_2(d)=\{c, d, e\}$, $\alpha_d(K_1, K_2)=3/5$.

The deviations can be ordered as follows:

$\alpha_a(K_1, K_2)=\alpha_h(K_1, K_2)=0$, $\alpha_b(K_1, K_2)=1/4$,

$\alpha_c(K_1, K_2)=\alpha_d(K_1, K_2)=3/5$, $\alpha_f(K_1, K_2)=\alpha_g(K_1, K_2)=2/3$, $\alpha_e(K_1, K_2)=1$.

The local similarities can be used to define a global similarity by

$$\alpha(K_1, K_2)=\frac{1}{|T_1\cup T_2|}\sum_{x\in T_1\cup T_2}\alpha_x(K_1, K_2)$$

or a global distance by $\delta(K_1, K_2)=1-\alpha(K_1, K_2)$.

Example. Let K_1 and K_2 be as above. Then we get $\alpha(K_1, K_2)=1/8\,.\,227/60=227/480$.

If we take the same set of terms, δ is a metric on the class of partitions, we get a pseudometric.

Example. $K_3=\{\{a, b\}, \{c, d\}, \{e, f, g\}\}$, $K_4=\{\{a, b\}, \{c, d\}, \{e, f\}, \{e, g\}, \{f, g\}\}$, $\alpha(K_3, K_4)=0$.

A term classification can be used for modifying a search request of a document retrieval system. If the search request is extended by adding all terms that are in the same class as the original terms of the request, then the concept of a pseudometric is adequate. The reason for this is that classifications with the distance zero modify the request the same way.

Let K_0 be the finest classification (which only has unit classes) and K an arbitrary classification, then we have

$\alpha(K_0, K)=$ number of classes of $K/|T|=1/$average cardinality of the classes of K.

This ratio can be considered as an indicator of how fine the classification K is. The overlap measure however delivers $\alpha(K_0, K)=$ const for each K, because $\alpha_x(K_0, K)=1$. This is another reason for preferring the Tanimoto measure.

In many test situations, a query set Q is given. It may happen that great differences between classifications deliver small differences in the retrieval result and vice versa. This effect can be explained by the fact that in the first case the differences between the classifications do not affect the neighbourhood of the terms in Q, while in the second case such differences do exist. In both cases Q is not representative for the test situation.

This is the reason why we plead for local similarity between classifications. Let T_0 be the set of terms in Q. Then we define

$$\alpha_{T_0}(K_1, K_2)=\frac{1}{|T_0|}\sum_{x\in T_0}\alpha_x(K_1, K_2) \text{ and } \delta_{T_0}(K_1, K_2)=1-\alpha_{T_0}(K_1, K_2),$$

where the local similarity $\alpha_x(K_1, K_2)$ is computed as above.

δ_{T_0} is a pseudometric on the classifications of the same set of terms T. If the distance is zero, this indicates that K_1 and K_2 are equivalent with res-

pect to the set of queries Q since K_1 and K_2 change the queries in the same way provided that they are used as mentioned above.

Experiments. The metric is incorporated in the FAKYR document retrieval system which has been implemented on an IBM 370/158 [3]. We have used a document collection consisting oi computer science abstracts of the ZDE (Zentralstelle Documentation Elektrotechnik). 1773 terms have been clustered applying a single link algorithm for several threshold values. The following examples show some local deviations between txe classifications computed with a threshold $T=0.55$ or $T=0.6$ respectively. For the single linkage clustering the metric is highly correlated with the distribution of the similarities among the terms. This delivers an efficient estimation for the metric.

E x a m p l e. Distribution of similarities among the terms and distance between the classifications of the neighboured thresholds

Interval of thresholds	Number of similarities within the interval	Distance between the classifications of the thresholds
0,35—0,40	399	0,312
0,40—0,45	674	0,496
0,45—0,50	140	0,092
0,50—0,55	257	0,272
0,55—0,60	449	0,302
0,60—0,65	75	0,036
0,65—0,70	40	0,119
0,70—0,75	718	0,255
0,75—0,80	46	0,019
0,80—0,85	24	0,012
0,85—0,90	13	0,008
0,90—0,95	16	0,006
0,95—1,0	1225	0,684

Correlation coefficient = 0.95, Rank correlation coefficient = 0.93.

Current experiments include:

1. Tests for checking the correlation with respect to other classifications (cliques e. g.).
2. Investigations into the interrelationship between metrics and retrieval results.
3. Construction of a bilingual associative thesaurus.

Conclusion. Our technique seems to be a good semi-automatic tool for detecting local deficiencies in classification systems. In big organizations it is necessary that classification permits growth and expansion to handle new information items. Therefore the importance of efficient software tools for supporting classification will grow even more rapidly in the near future.

REFERENCES

1. M. R. Anderberg. Cluster Analysis for Application. New York, 1973.
2. J. G. Augustson, J. Minker. An Analysis of some Graph Theoretical Cluster Techniques. J. Assoc. Comput. Mach., 17, 1970, 571-588.

3. M. Bock, H. L. Hausen, E. Konrad, H. Zuse. FAKYR — An Online Information Retrieval System. Proceedings of the 1975 Conference on Information Sciences and Systems, Baltimore, 1975.
4. P. Bollmann, E. Konrad. Automatic Association Methods in the Construction of Interlingual Thesauri. ASLIB Conference EURIM 2. Amsterdam, March 23-25, 1976.
5. D. D. Gotlieb, S. Kumar. Semantic Clustering of Index Terms. *J. Assoc. Comput. Mach.*, **15**, 1968, 493—513.
6. J. A. Hartigan. Representation of Similarity Matrices by Trees. *J. Amer Statist. Assoc.*, **62**, 1967, 1140—1158.
7. G. Salton. Automatic Information Organization and Retrieval. New York, 1968.
8. K. Sparck-Jones. Automatic Keyword Classification and Information Retrieval. London, 1971.

Technische Universität Berlin
Berlin DDR

Received 11. 9. 1977