Using Cluster Analysis for Processing English Texts

Yong Kui Zhang
Department of Computer Sience
Shanxi University, Taiyuan 030006, P.R.China
James R Cowie
Department of Computing Science
University of Stirling, Stirling FK9 4LA Scotland U.K.

Abstract

Some simple processing techniques have allowed the application of a standard software package to the areas of Text Analysis and Natural Language Processing. The techniques, in general, rely only on word count and word co-occurrence information which is derived directly from the character representation of the text; no additional semantic or grammatical information is used. The analysis technique used, Cluster Analysis, groups objects together based on measures of similarity or difference between the values of a set of variables representing each object. A variety of simple text manipulation processes were used to produce appropriate input formats for the statistical package.

Three different text processing tasks were chosen. The first was to categorize a group of short, one paragraph, newspaper articles. The second was to disambiguate the word 'bank' occurring in a set of 400 sentences and in a set of 200 paragraphs based solely on the other words occurring in the sentence or paragraph. The third task was an attempt to replicate the mammalian classification system based on an analysis of short texts describing 426 animals. In general the technique gives interesting results and it would seem likely that it could usefully be used in many subject areas which give rise to homogeneous sets of text.

1 Introduction

Good results have been achieved recently in the areas of Text Analysis and Natural Language processing using techniques largely dependent on word count and word co-occurrence information [1, 2]. The techniques, in general, rely only on information which is derived directly from the character representation of the text; no additional semantic or grammatical information is used. These tasks have been performed using special purpose programs. Given the success of these methods it seemed natural to ask whether it is possible to perform equivalent tasks using a standard statistical software package. This would minimize the complexity of any programming effort required and would allow a more exploratory approach to be adopted. SPSS-X [3] (Statistical Package for Social Sciences - eXtended) was chosen for the work. It is generally available and is capable of handling large numbers of variables. The analysis technique used, Cluster Analysis, groups objects together based on measures of similarity or difference between the values of a set of variables representing each object. A variety of simple text manipulation processes were used to produce appropriate input formats for the statistical package.

Three different text processing tasks were chosen. The first was to categorize a group of short, one paragraph, newspaper articles. This task is a traditional use of cluster analysis in Information Retrieval systems for partitioning texts into appropriate groups [4]. The second was to disambiguate the word 'bank' occurring in a set of 400 sentences and in a set of 200 paragraphs based solely on the other words occurring in the sentence or paragraph. In this case an alternative technique had been recently described [1] with which performance could be compared. The third task was an attempt to replicate the mammalian classification system based on an analysis of short texts describing 426 animals. This exercise was prompted by the availability of a machine readable version of the texts. Highly accurate results were not expected as the textual descriptions were not particularly complete.

Final measures of performance in each of the tasks are shown in Table 1.

Table 1: Test Results

Object Type	Topic	Number	Correct	Ratio	Objective
Articles	Home News	82	63	77 %	Categorizing
Paragraphs	Bank	200	174	87 %	Disambiguation
Sentences	Bank	400	364	90 %	Disambiguation
Description	Mammals	426	318	75 %	Classification

2 Cluster Analysis

The techniques of cluster analysis are useful tools for data analysis in many diverse fields: psychology, zoology, biology, botany, sociology, artificial intelligence and information retrieval. The problem which these techniques attempt to solve was defined by Everitt [5] as follows:

> 'Given a sample of N objects or individuals, each of which is measured on each of p variables, devise a classification scheme for grouping the objects into g classes.'

The result of a cluster analysis is a number of groups, clusters, or classes. In general the initial raw data consists of Nxp matrix of measurements. In this paper the entities which are to be clustered are representations of English texts: articles, paragraphs, sentences and descriptions. The variables measured are the specific words which occur in these texts.

There are five basic cluster analysis techniques [5]. Of these the hierarchical techniques are the ones normally used for computer implementation. The method used by SPSS-X is an aglommerative hierarchical clustering technique.

The basic process with all agglomerative methods is similar:

1. Compute (or input) the proximities between the initial clusters (the individual cases)
2. Combine the two nearest clusters to form a new cluster
3. Recompute the proximities between existing clusters and the new cluster
4. Return to the second step until all cases have been combined in one cluster

At any particular stage the methods fuse clusters or groups of clusters which are nearest (or most similar). Differences between methods arise because of the different ways of defining distance (or similarity) between a cluster and a group containing several clusters, or between two groups of clusters [6, 7].

The majority of clustering techniques begin with the calculation of a matrix of similarity, or dissimilarity, or distance measures between clusters. Similarity measures increase with greater similarity; dissimilarity and distance measures decrease. A similarity coefficient measures the relationship between two items, given the values of a set of p variates common to both. Similarity coefficients take values in the range 0 to 1. Here the variates are of the 'presence' and 'absence' type which may be arranged in the familiar two-way association table shown in Table 2.

The most useful binary measures for our text analysis problem proved to be the **Ochiai Similarity Measure**, the **Binary Shape Difference** and the **Binary Squared Euclidean Distance**. Three distance measures were also used: the **Euclidean Metric**, the **City-block**, or **Manhattan Distance** and the **Distance in an Absolute Power Metric**. All these measures are fully described in the SPSS-X manual.

Table 2: Two-way association table for two items

	Item 2 characteristics	
	Presence	Absence
Item 1 characteristics		
Presence	a	b
Absence	c	d

3 Automatic Text Analysis

Before using clustering techniques for processing a set of texts, it is necessary to perform a series of text analysis processes which generate from the actual text a document representation which is appropriate for use in the clustering process. These processes consist of:

1. re-organizing texts
2. pre-processing texts
3. generating a text representation

The starting point of the text analysis process is an original form of text: printed or electronic. The text re-organization involves converting the original form into a fixed one which can be used by the computer. For example, our internal forms in two cases were derived from CD-ROM text: an electronic newspaper on CD-ROM and an electronic encyclopedia. In the case of the mammal texts a printed version of the text had been scanned using an **Optical Character Reader**. It had then been subjected to intensive correction.

Specific sections of the text were then extracted using a selection program For example, sentences or a paragraphs including the word 'bank(s)' were pulled from longer texts extracted from *Grolier's Encylopedia* [8].

Once the object texts have been generated they are pre-processed. An object text is a connected text, it can be handled as a linear string of characters, and broken up into words of varying length which can then be processed. Pre-processing the texts included removing all punctuation marks and other non-alphabet characters, and sometimes converting upper case letters into lower case.

The representation used for processing is not the complete object text. A text representation is produced from the complete object text. Luhn [9] assumed that frequency data can be used to extract words and sentences to represent a document. Schultz and Luhn [10] given a plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order. This indicated that the words exceeding the upper cut-off were considered to be common and those below the

Table 3: Suffixes

-ed	-er	-ess	-est	-ful
-ical	-ing	-ish		
-less	-like	-ly		
-ment	-ness	-ous	-s	

lower cut-off rare, neither group contributing significantly to the content of the article.

Frequency counts were used for our four sets of texts to determine which words are sufficiently significant to represent or characterize the texts. The process of generating a text representation then consisted of two parts:

1. removal of high frequency words
2. suffix stripping

The removal of high frequency words was done by comparing the input text with a 'stop list' of words which are to be removed. The list for the work described here was adapted from the first two hundred words of the Birmingham Corpus. The advantages of this process are not only that non-significant words are removed and will therefore not interfere during classification, but also that the size of the total document file can be reduced by about 30 per cent.

Suffix stripping is more complicated [4]. This process was done in a relatively crude manner by comparing the ending of every word in the input text with a suffix list and modifying those words which matched.

The suffixes in Table 3 were processed.

The final representation produced for each text was a list of words which were within the input text and had been processed with the above methods. Every word presents itself just once in the text representation. So after this processing the size of the total document file can be reduced by between 30 and 70 per cent. No information on frequency of occurrence is kept.

4 Automatic classification

We used the SPSS-X System to cluster each set of texts. The process has three phases:

1. generating an SPSS-X input file
2. computing a proximity matrix
3. performing cluster analysis

Creating the SPSS-X file is question of selecting the particular clustering options to be used and associating the commands for these options with the appropriate data-files. The variables for all the clustering techniques dealt with here are the words of the document representations which need to be classified.

The procedure of generating a wordlist (the variables to be measured) included counting word frequency and removal of lower frequency words. Word frequency here means the number of documents a word occurred in. That is, frequency 1 indicates that a word occurred only in one document, frequency 2 means that a word occurred in two documents, and so on. The removal of lower frequency words depends on the number of variables for which similarity coefficients are computed. Since storage requirements increase rapidly with the number of cases and the number of variables, it is necessary to choose moderate-sized data sets within the workspace limitations of the computer system. For example, we computed a proximity matrix for 426 cases (for mammal descriptions) and with 940 variables (words) in a 4000K-byte workspace using SPSS-X Release 3.0 for HP-UX.

After generating a wordlist – the variable list to be used as input to the clustering method, the next stage is to generate SPSS-X input file containing the raw data and the necessary commands to control SPSS-X. This process was carried out automatically by comparing each input text representation with the wordlist.

There are 37 proximity measures provided by SPSS-X.. Which similarity measure should be used, since different measures may lead to different results? There can be no absolute answer to such a question [12]. In many cases when using dichotomous variables the major decision to be made is which similarity coefficient should be used to measure similarity between entities. In general, the major problem is that of the otherwise of negative matches. Different binary measures emphasize different aspects of relation between sets of binary values. According to the structure of the raw data matrices we choose binary measures–Ochiai Similarity Measure, or Binary Shape Difference, or Binary Squared Euclidean Distance. These gave good results.

After choosing the proximity measures, the next stage is to choose the clustering technique or techniques. No one technique can be judged to be 'best' in all circumstances. In general several techniques should be used, as this should help to prevent misleading solutions being accepted. SPSS-X provides 7 methods for linking cases in the agglomeration process. BAVERAGE or COMPLETE methods both proved to be satisfactory. BAVERAGE is an average linkage between groups. COMPLETE is complete linkage, or furthest neighbour.

5 Performance in the Three Tasks

5.1 Categorizing Articles

Categorizing articles is helpful in information access in complex, poorly structured information spaces. As a test, we randomly selected 82 articles about home news from a British newspaper (*The Guardian*) on CD-ROM. The size of these articles was about 44 thousand characters. The number of different words included in the articles was 3022. After automatically analyzing these texts we selected 200 words to be used as the variables on the 82 objects. The proximity measure Bseuclid was used followed a Block measure.

The actual task of categorizing is a difficult one for humans. We attempted several groupings based on simple either/or questions of the text. For example, did the article describe some violent action? The texts gave rise to six clusters; one large containing 50 texts, 36 of which described violent actions. The remaining 5 smaller clusters contained mostly non-violent texts, 27 out of 32. The length of the texts and their wide range of subject matter meant that there was insufficient overlap of common words to allow the clustering process to operate really satisfactory. Three texts grouped together because they shared only one word in common 'yesterday'.

5.2 Word sense disambiguation

Clustering techniques can be used to aid word sense disambiguation. Firstly, we selected randomly 200 paragraphs including word BANK(S) from *The Electronic Encyclopedia (A 20-Volume Encyclopedia on CD-ROM)*. Then we selected randomly 400 sentences including word 'BANK(S)' from above paragraphs. The size of these paragraphs was about 135 thousand characters. The number of different words used was 2533. 740 words were selected as variables. An Ochiai proximity measure was used.

The result was that 174 of the paragraphs clustered correctly according to whether the river or money sensed of the word 'bank' was intended.

400 individual sentences were then extracted from the paragraphs. The number of the total different words included in the 400 sentences was 2035. 260 words were selected as the variables of these sentences. In this case a Bshape measure was used. The performance of the method was even higher with 364 sentences being clustered correctly.

5.3 Mammals classification

The book *The Macdonald Encyclopedia of Mammals* [11] was processed. This book describes 426 species of mammals, from minute shrews to gigantic whales, which are to be found throughout the world. Each entry contains the mammal's classification, description, distribution and habitat, as well as details on

behavior, feeding habits and reproduction. The size of this book is about 580 thousand characters. The number of the total different words used in this book was 6407. Each mammal's description, habitat, and behavior was extracted from this book as its representation text. The size of the 426 texts is about 465 thousand characters. The number of the total different words used in these texts is 3849. After automatically analyzing these texts, 940 words were selected as the keywords. The proximity measure Ochiai was used followed by a Block measure.

The results of this analysis were surprisingly good. Seventy five percent of the descriptions were classified into the correct mammalian order. Given that the descriptions are not full (being supplemented by pictures) and that we could easily be clustering all mammals with red fur, wintry habitats or any of a wide range of corresponding characteristics. Some of these, of course, may be related to the mammalian order. In the case of certain well distinguished orders of mammal - bats (chiroptera) and all the sea living mammals (cetacea) the method was completely accurate. Part of the dendrogram produced by SPSS-X is given in the appendix which illustrates one type of output produced by the system. Each object text is named by the mammalian order and the text number.

6 Conclusions

Some simple processing techniques have allowed the application of a standard software package to a variety of applications. In the case of disambiguation the performance is as good as any other technique. It is our intention to try to compare the results of other disambiguation techniques with this one. In particular it would be interesting to observe whether the methods fail on different texts.

For the clustering method to succeed it is necessary for a good overlap to exist in the words occurring in the middle frequencies of the text. Our categorization of newspaper articles has highlighted this problem and we now intend to establish some guidelines relating vocabulary size and text length.

In the case of the mammals it may be possible to improve the results by limiting the vocabulary using semantic techniques. For example by using dictionary definitions from a machine readable dictionary to select only words naming parts of the animal a much more precise definition of similarity may be produced.

In general the technique gives interesting results and it would seem likely that it could usefully be used in many subject areas which give rise to homogeneous sets of text.

References

[1] J A GUTHRIE, et al,*Subject-Dependent Co-occurrence and Word Sense*

Disambiguation, Proceedings of the 29th ACL Annual Meeting, June, (1991)

[2] P F BROWN, et al, *A statistical approach to machine translation*, Technical Report RC 14773, IBM Research Division, T J Watson Research Center (1989)

[3] *SPSS-X User's Guide*, 3rd Edition, SPSS Inc.

[4] C J VAN RIJSBERGEN, *Information Retrieval*, 2nd ed., Butterworths, London (1979)

[5] B EVERITT, *Cluster Analysis*, Heinemann Educational Books Ltd, London (1974)

[6] G N LANCE, W T WILLIAMS, *A general theory of classificatory sorting strategies: 1. hierarchical systems*, Comp. J, **9**, 373-380 (1967)

[7] D WISHART, *An algorithm for hierarchical classifications*, Biometrics, **25**, 165-170 (1969)

[8] *The Electronic Encyclopedia - User's Guidei*, Grolier Electronic Publishing, Inc., Danbury, CT (1986)

[9] H P LUHN, *The automatic creation of literature abstracts*, IBM Journal of Research and Development, **2**, 159-165 (1958)

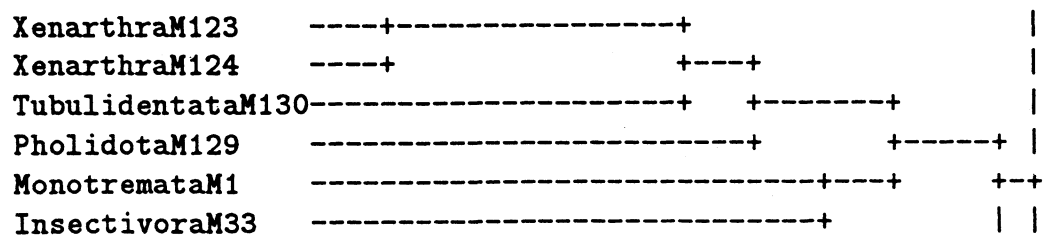
[10] C K SCHULTZ, H P LUHN, *Pioneer of Information Science-Selected Words*, Macmillan, London (1968)

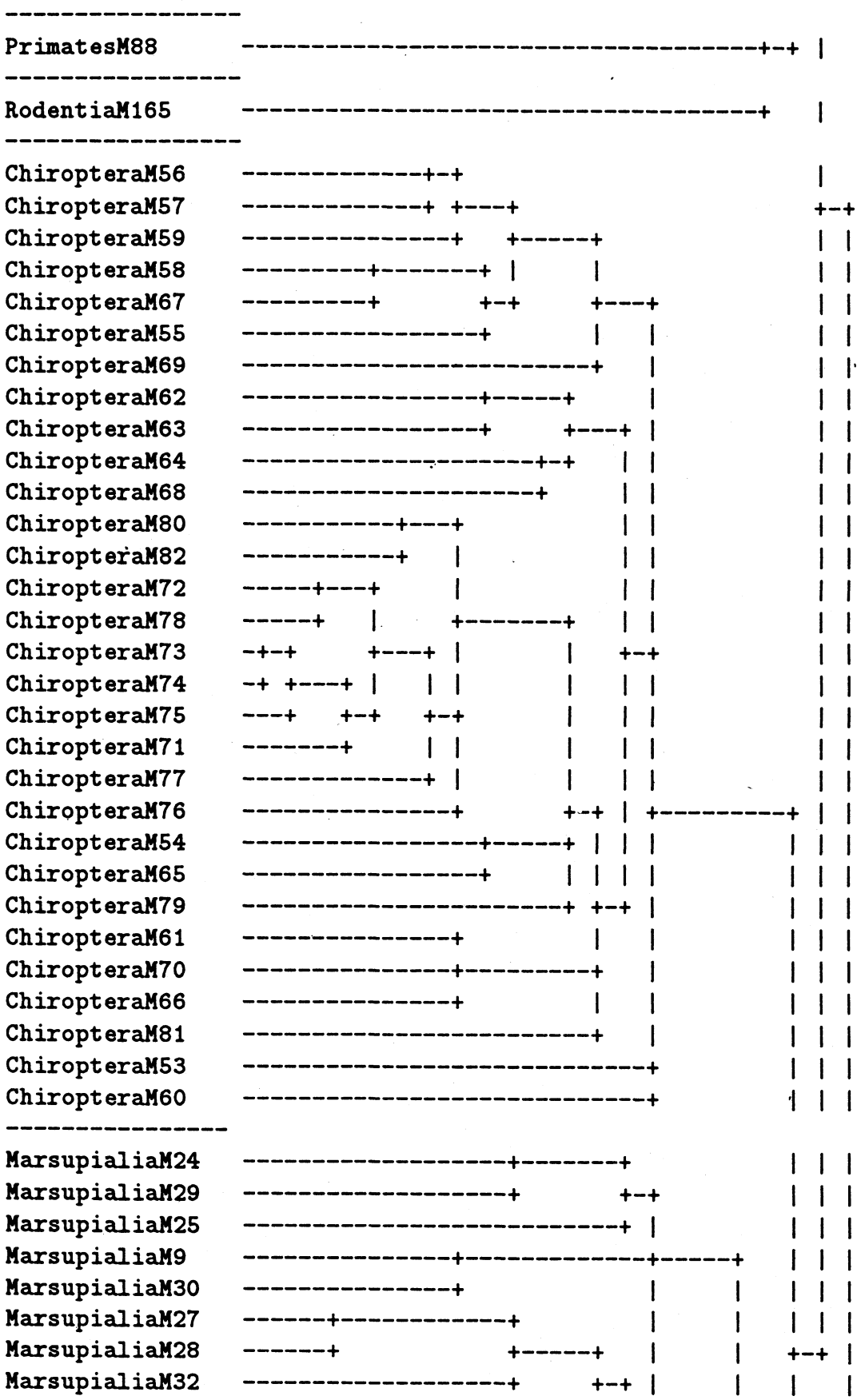
[11] L BOITANI, S BARTOLI, *The Macdonald Encyclopedia of Mammals*, Macdonald, London (1983)

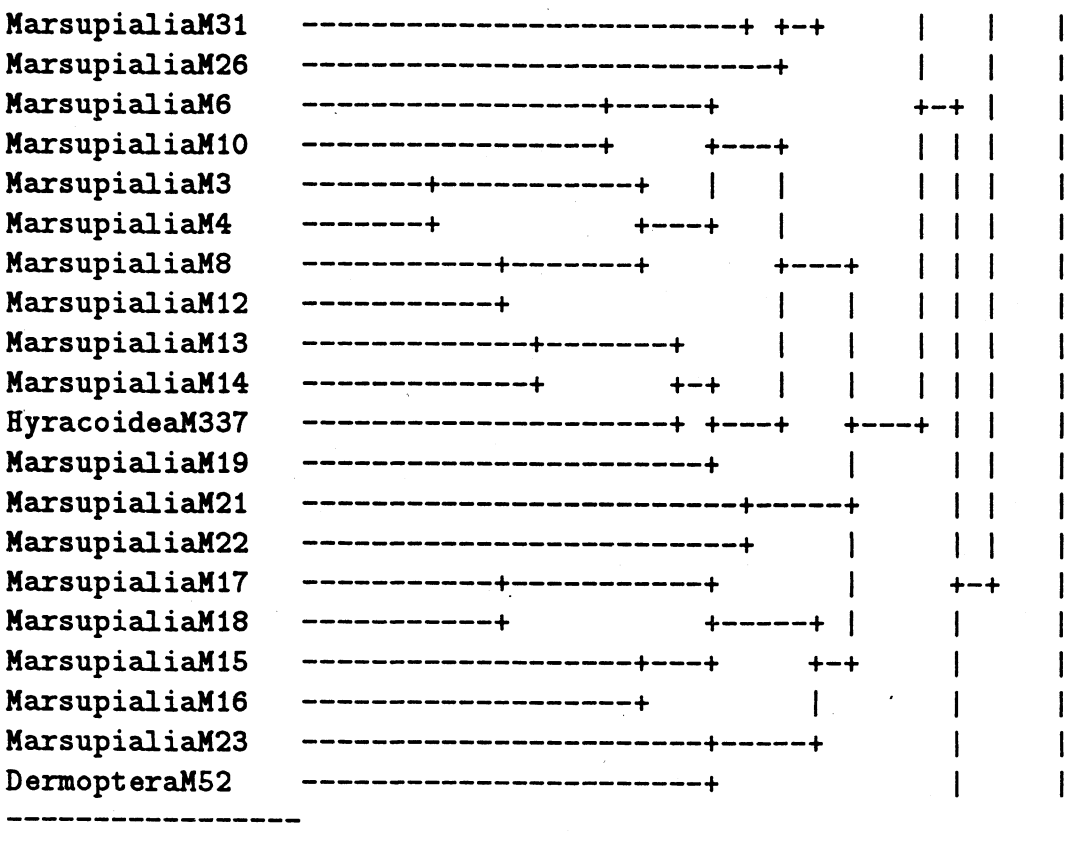
[12] R R SOKAL, P H A SNEATH, *Principles of Numerical Taxonomy*, W H Freeman and Company, San Francisco and London (1963)

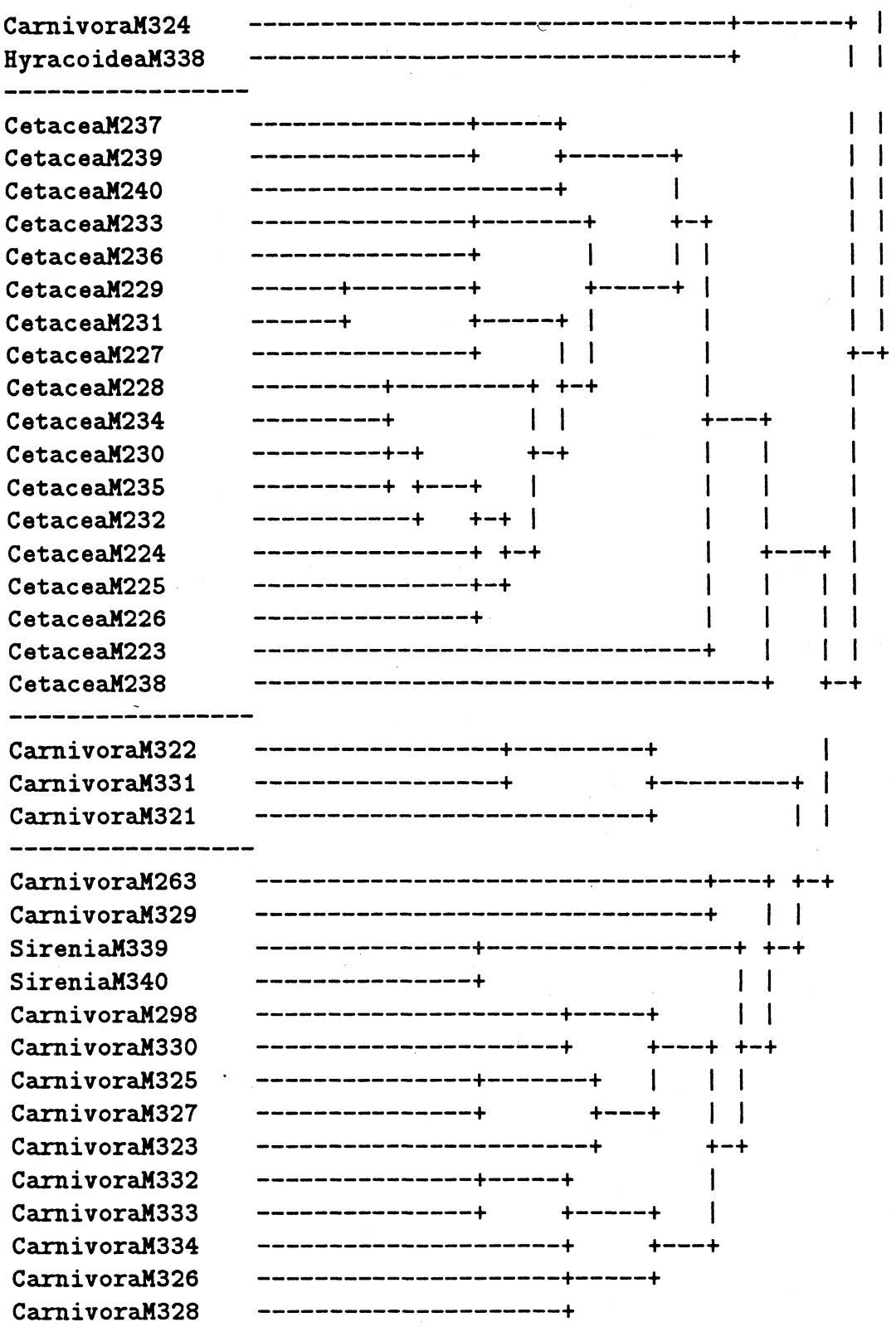
Appendix

Part of Mammal Classification Dendrogram




PrimatesM88
RodentiaM165
ChiropteraM56
ChiropteraM57
ChiropteraM59
ChiropteraM58
ChiropteraM67
ChiropteraM55
ChiropteraM69
ChiropteraM62
ChiropteraM63
ChiropteraM64
ChiropteraM68
ChiropteraM80
ChiropteraM82
ChiropteraM72
ChiropteraM78
ChiropteraM73
ChiropteraM74
ChiropteraM75
ChiropteraM71
ChiropteraM77
ChiropteraM76
ChiropteraM54
ChiropteraM65
ChiropteraM79
ChiropteraM61
ChiropteraM70
ChiropteraM66
ChiropteraM81
ChiropteraM53
ChiropteraM60
MarsupialiaM24
MarsupialiaM29
MarsupialiaM25
MarsupialiaM9
MarsupialiaM30
MarsupialiaM27
MarsupialiaM28
MarsupialiaM32


MarsupialiaM31
MarsupialiaM26
MarsupialiaM6
MarsupialiaM10
MarsupialiaM3
MarsupialiaM4
MarsupialiaM8
MarsupialiaM12
MarsupialiaM13
MarsupialiaM14
HyracoideaM337
MarsupialiaM19
MarsupialiaM21
MarsupialiaM22
MarsupialiaM17
MarsupialiaM18
MarsupialiaM15
MarsupialiaM16
MarsupialiaM23
DermopteraM52
InsectivoraM50
InsectivoraM51
InsectivoraM38
InsectivoraM45
InsectivoraM49
InsectivoraM37
MonotremataM2
MarsupialiaM5
InsectivoraM34
InsectivoraM35
MacroscelidiaM41
MacroscelidiaM42
MarsupialiaM11
MarsupialiaM20
InsectivoraM36
InsectivoraM40
InsectivoraM44
InsectivoraM46
InsectivoraM43
InsectivoraM47
InsectivoraM48
InsectivoraM39


CarnivoraM324
HyracoideaM338
CetaceaM237
CetaceaM239
CetaceaM240
CetaceaM233
CetaceaM236
CetaceaM229
CetaceaM231
CetaceaM227
CetaceaM228
CetaceaM234
CetaceaM230
CetaceaM235
CetaceaM232
CetaceaM224
CetaceaM225
CetaceaM226
CetaceaM223
CetaceaM238
CarnivoraM322
CarnivoraM331
CarnivoraM321
CarnivoraM263
CarnivoraM329
SireniaM339
SireniaM340
CarnivoraM298
CarnivoraM330
CarnivoraM325
CarnivoraM327
CarnivoraM323
CarnivoraM332
CarnivoraM333
CarnivoraM334
CarnivoraM326
CarnivoraM328