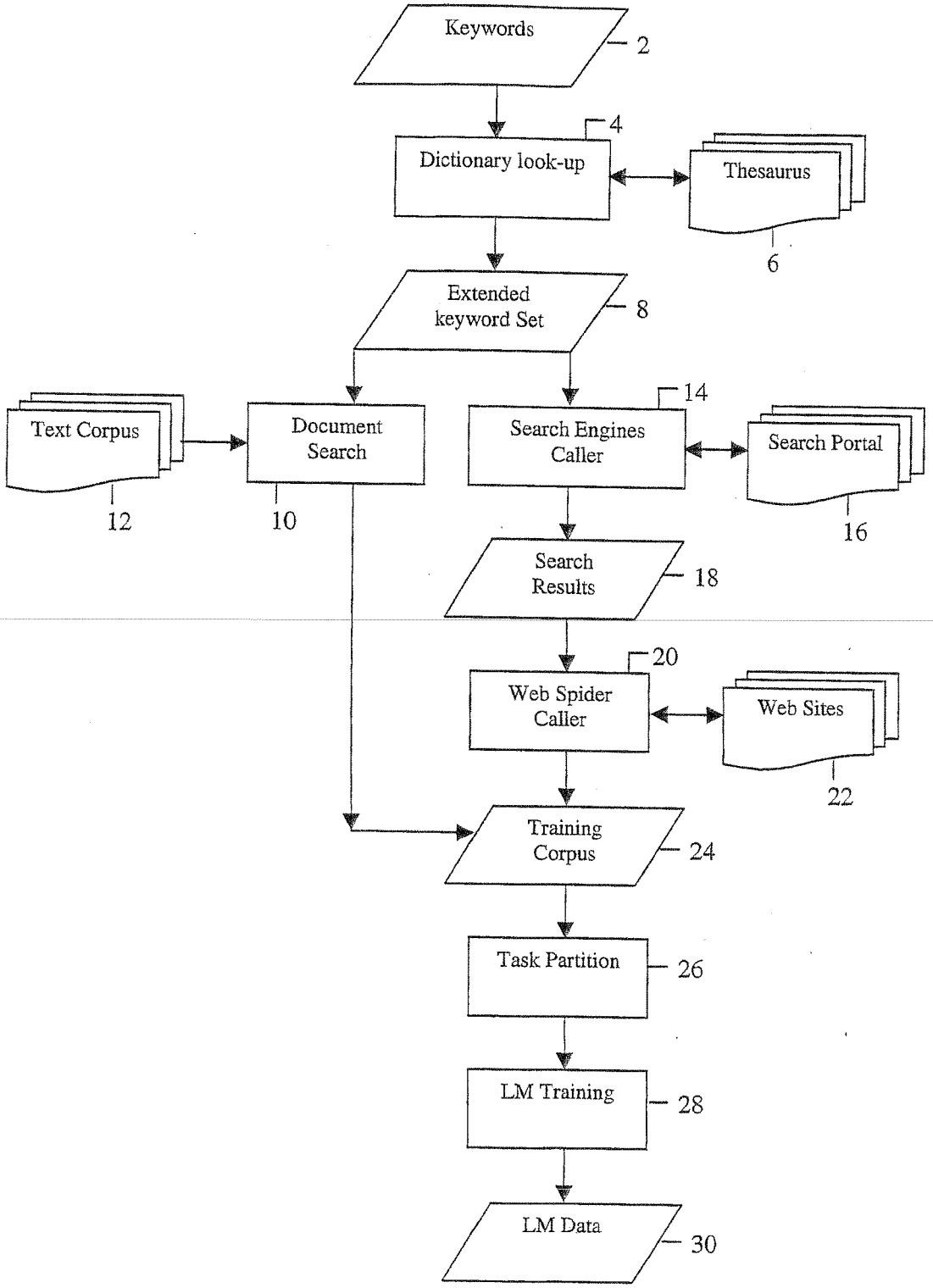

Keywords
2
Dictionary look-up
4
Thesaurus
6
Extended keyword Set
8
Text Corpus
12
Document Search
10
Search Engines Caller
14
Search Portal
16
Search Results
18
Web Spider Caller
20
Web Sites
22
Training Corpus
24
Task Partition
26
LM Training
28
LM Data
30

Figure 1

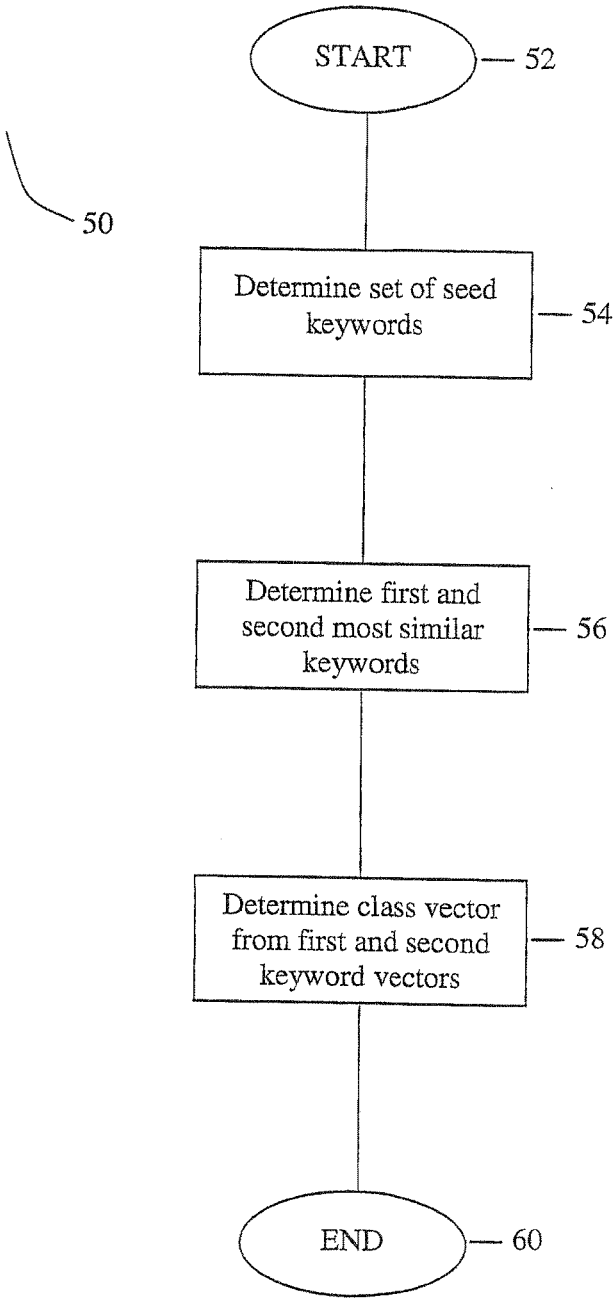

START
52
50
Determine set of seed keywords
54
Determine first and second most similar keywords
56
Determine class vector from first and second keyword vectors
58
END
60

Figure 2

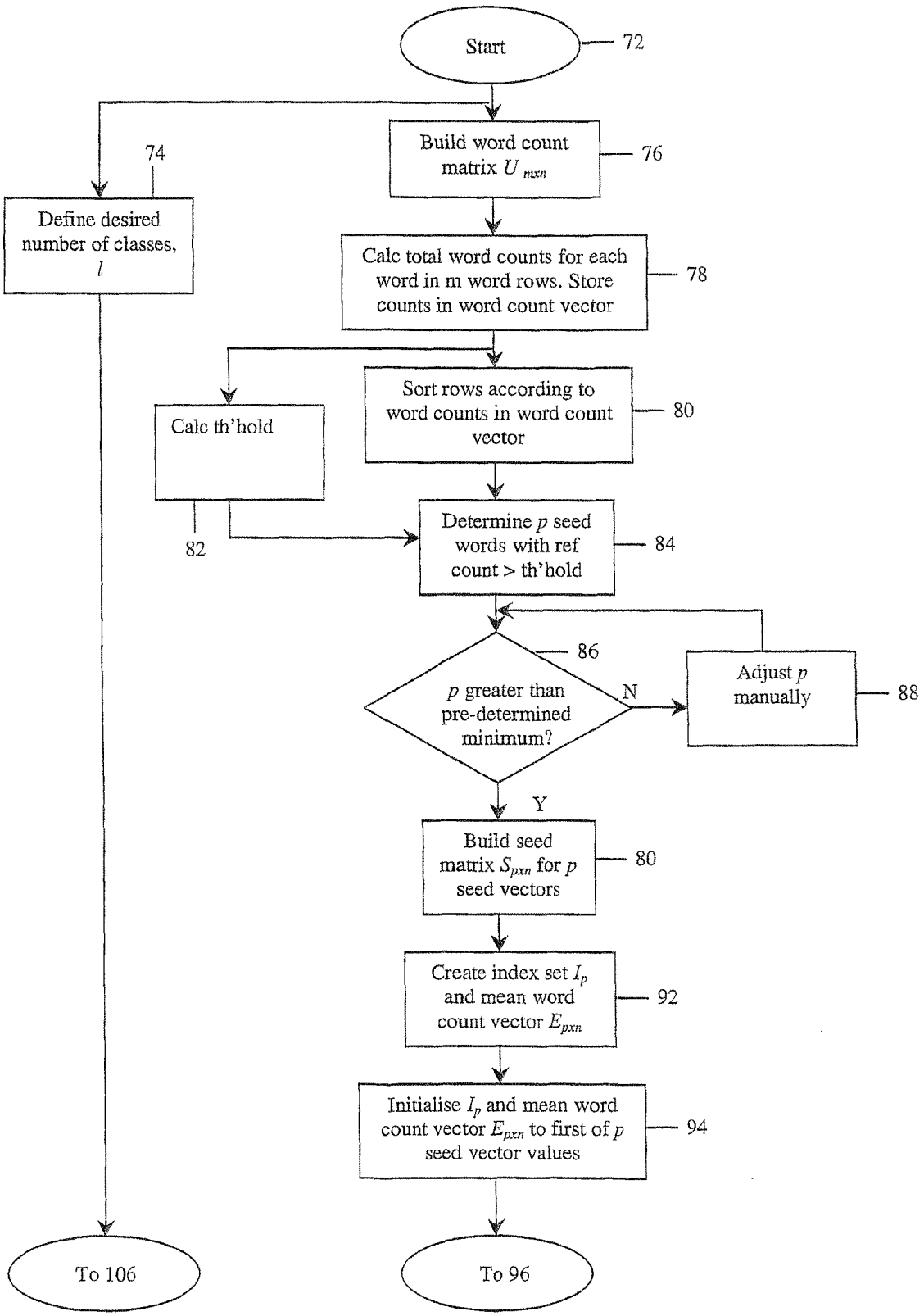

Start
72
Build word count matrix $U_{mxn}$
76
74
Define desired number of classes, $l$
Calc total word counts for each word in m word rows. Store counts in word count vector
78
Sort rows according to word counts in word count vector
80
Calc th'hold
82
Determine $p$ seed words with ref count > th'hold
84
86
$p$ greater than pre-determined minimum?
N
Adjust $p$ manually
88
Y
Build seed matrix $S_{pxn}$ for $p$ seed vectors
80
Create index set $I_p$ and mean word count vector $E_{pxn}$
92
Initialise $I_p$ and mean word count vector $E_{pxn}$ to first of $p$ seed vector values
94
To 106
To 96

Figure 3a

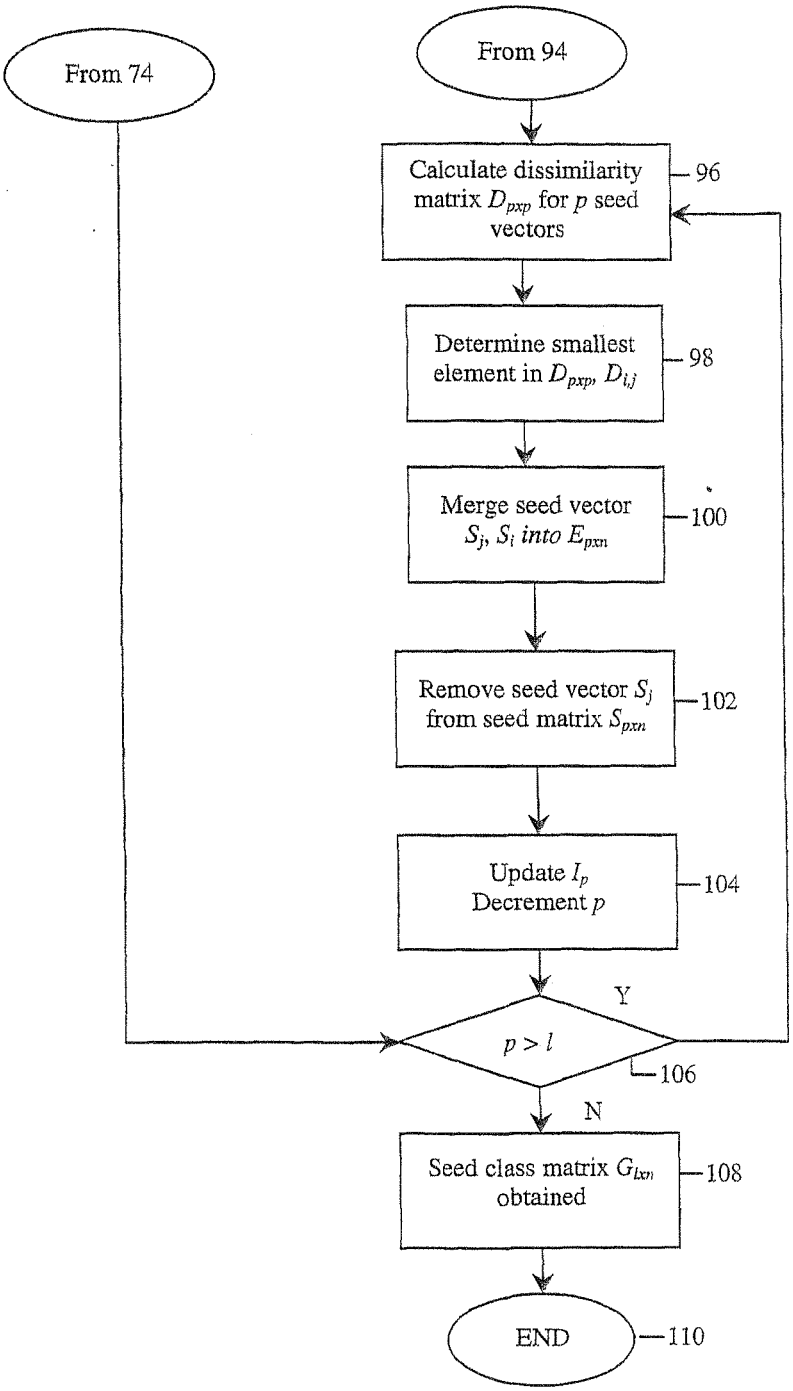
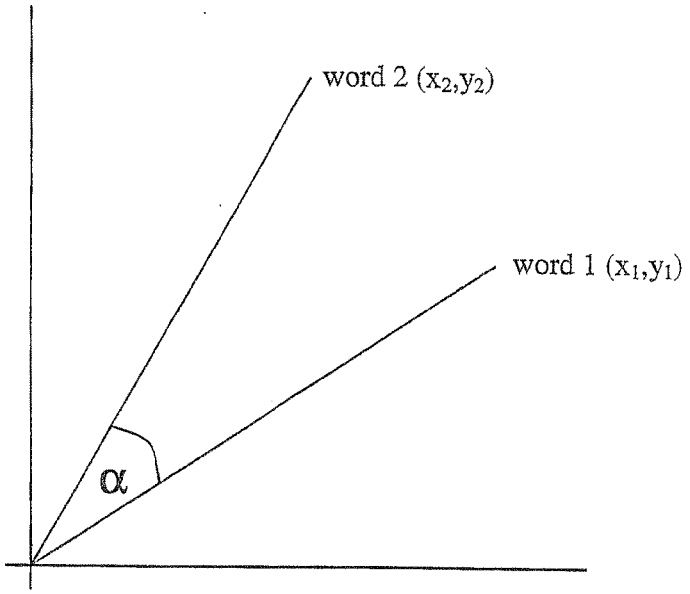

From 74
From 94
Calculate dissimilarity matrix $D_{pxp}$ for $p$ seed vectors
96
Determine smallest element in $D_{pxp}$, $D_{i,j}$
98
Merge seed vector $S_j$, $S_i$ into $E_{pxn}$
100
Remove seed vector $S_j$ from seed matrix $S_{pxn}$
102
Update $I_p$
Decrement $p$
104
Y
$p > l$
106
N
Seed class matrix $G_{lxn}$ obtained
108
END
110

Figure 3b




$$\text{Cos } \alpha = \frac{x_1 \cdot x_2 + y_1 \cdot y_2}{\sqrt{(x_1 + x_2)^2 (y_1 + y_2)^2}} \qquad (1)$$

Figure 3c

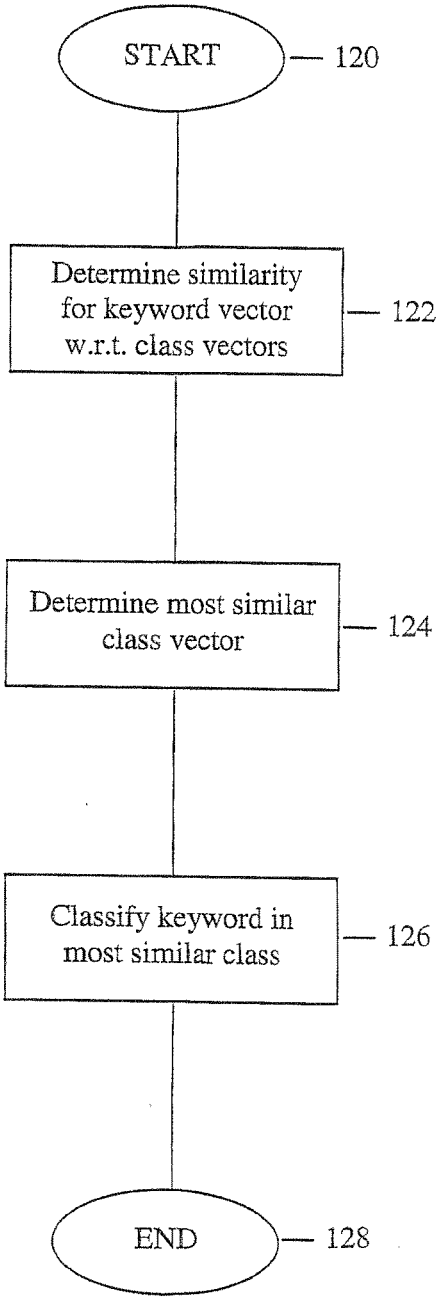

START
120
Determine similarity for keyword vector w.r.t. class vectors
122
Determine most similar class vector
124
Classify keyword in most similar class
126
END
128

Figure 4

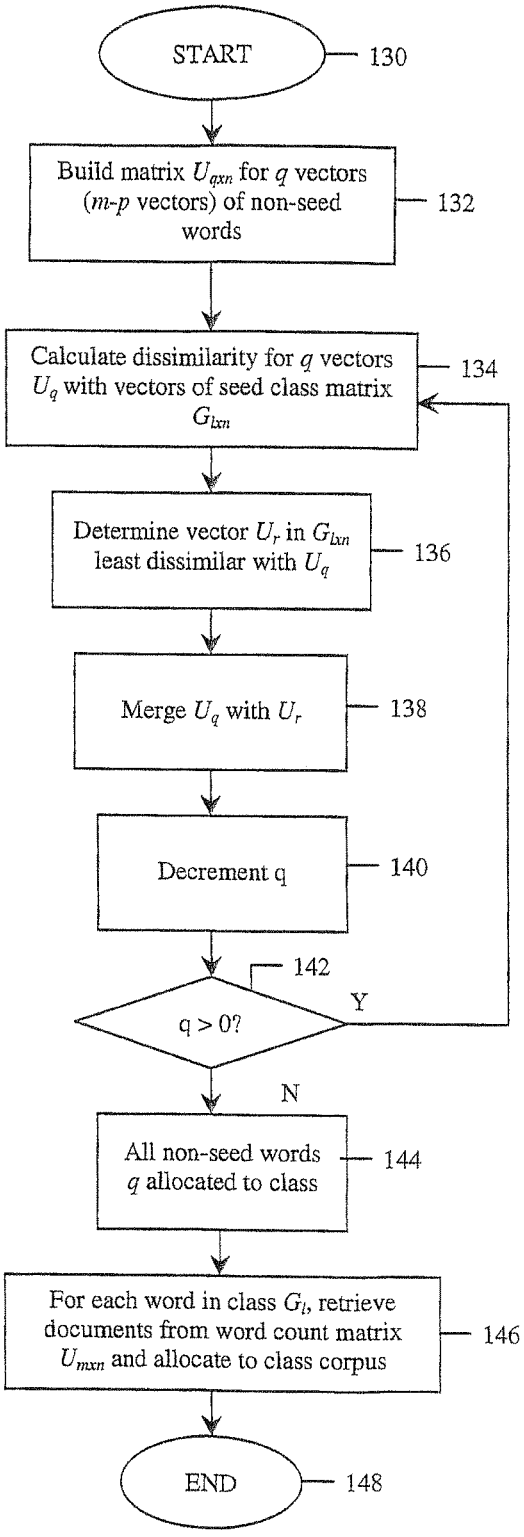

START
130
Build matrix $U_{qxn}$ for $q$ vectors ($m$-$p$ vectors) of non-seed words
132
Calculate dissimilarity for $q$ vectors $U_q$ with vectors of seed class matrix $G_{lxn}$
134
Determine vector $U_r$ in $G_{lxn}$ least dissimilar with $U_q$
136
Merge $U_q$ with $U_r$
138
Decrement q
140
142
q > 0?
Y
N
All non-seed words $q$ allocated to class
144
For each word in class $G_l$, retrieve documents from word count matrix $U_{mxn}$ and allocate to class corpus
146
END
148

Figure 5

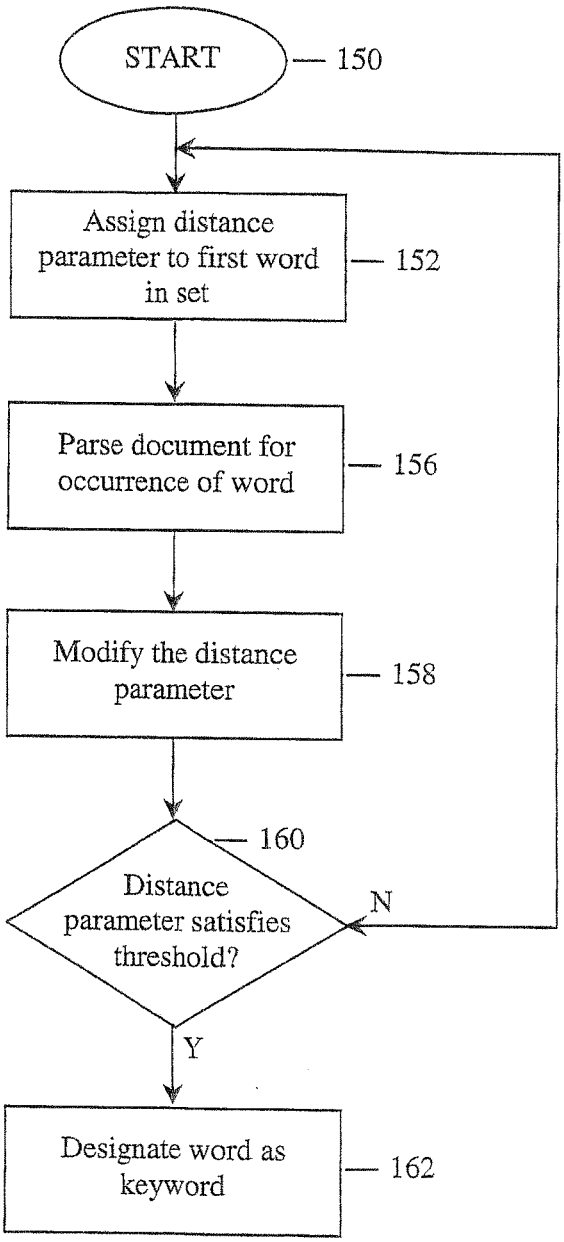

START
150
Assign distance parameter to first word in set
152
Parse document for occurrence of word
156
Modify the distance parameter
158
160
Distance parameter satisfies threshold?
N
Y
Designate word as keyword
162

Figure 6

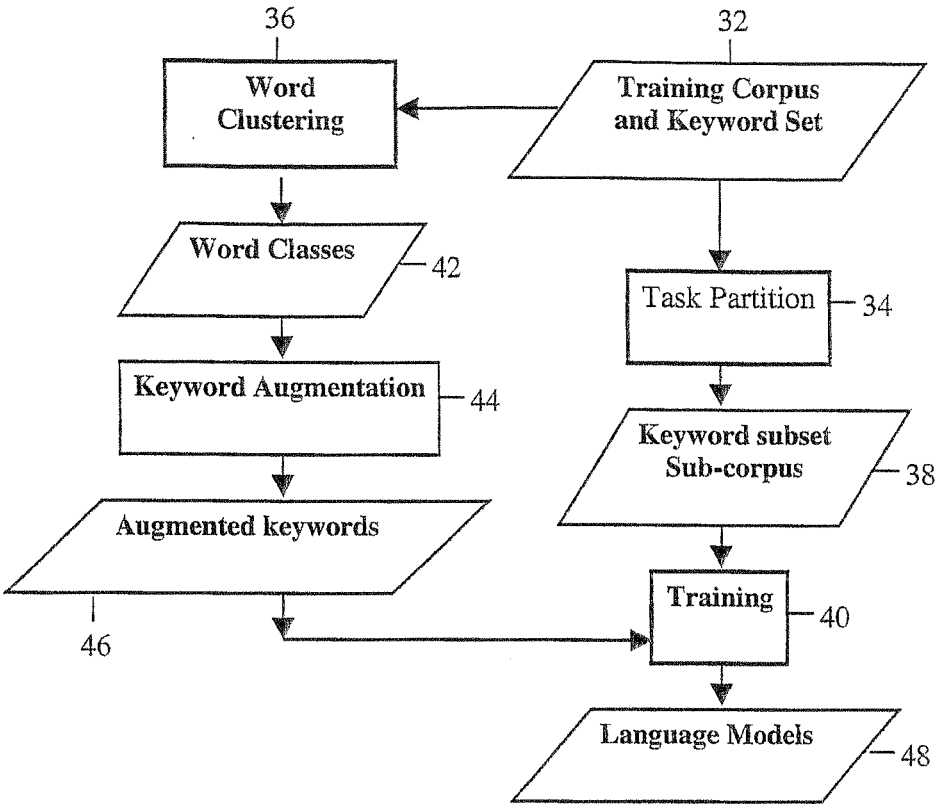

36
32
Word Clustering
Training Corpus and Keyword Set
Word Classes
42
Task Partition
34
Keyword Augmentation
44
Keyword subset Sub-corpus
38
Augmented keywords
46
Training
40
Language Models
48

Figure 7

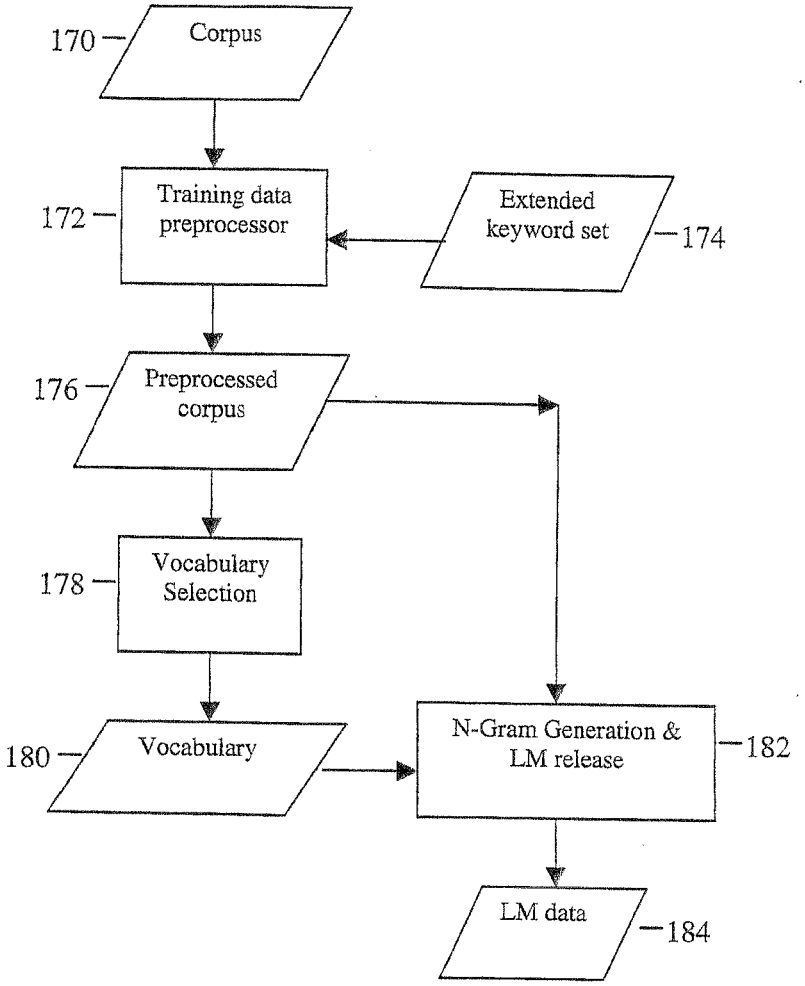

170
Corpus
172
Training data preprocessor
Extended keyword set
174
176
Preprocessed corpus
178
Vocabulary Selection
180
Vocabulary
N-Gram Generation & LM release
182
LM data
184

Figure 8

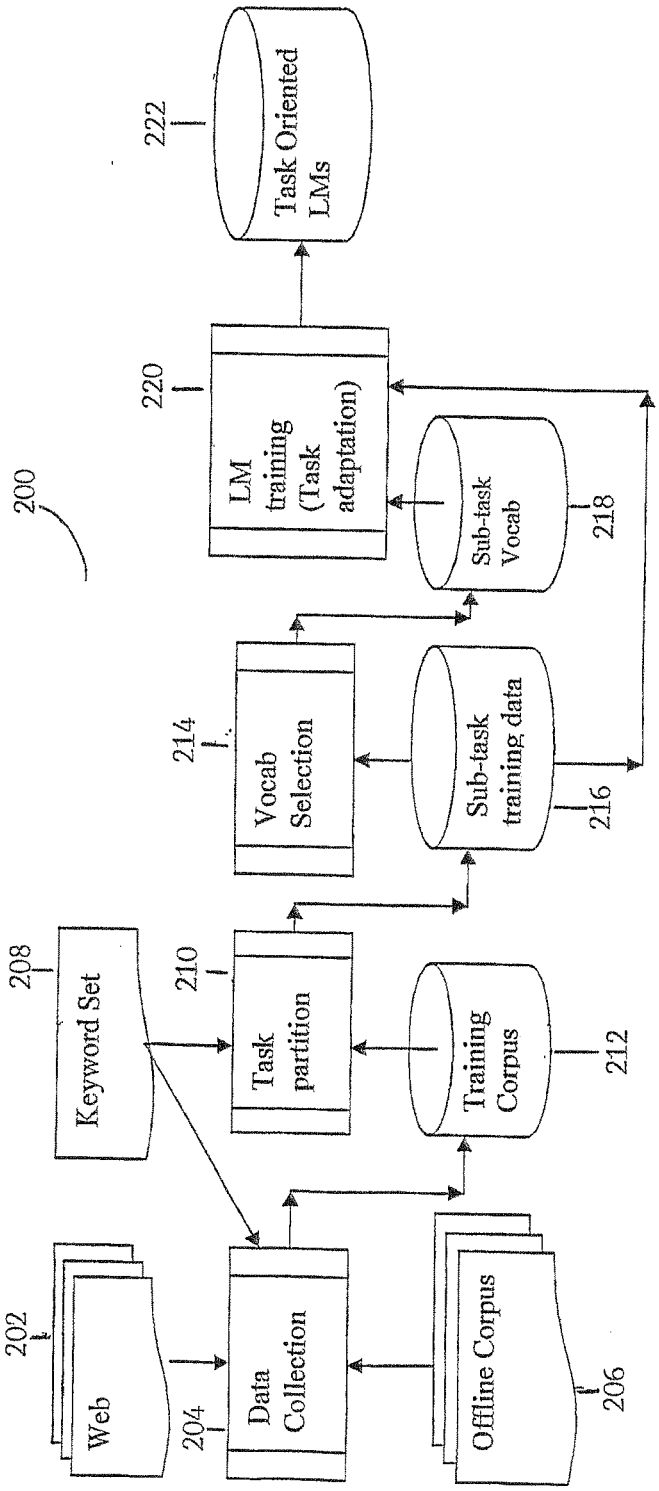

200
202
Web
204
Data Collection
206
Offline Corpus
208
Keyword Set
210
Task partition
212
Training Corpus
214
Vocab Selection
216
Sub-task training data
218
Sub-task Vocab
220
LM training (Task adaptation)
222
Task Oriented LMs

Figure 9