FIG. 1

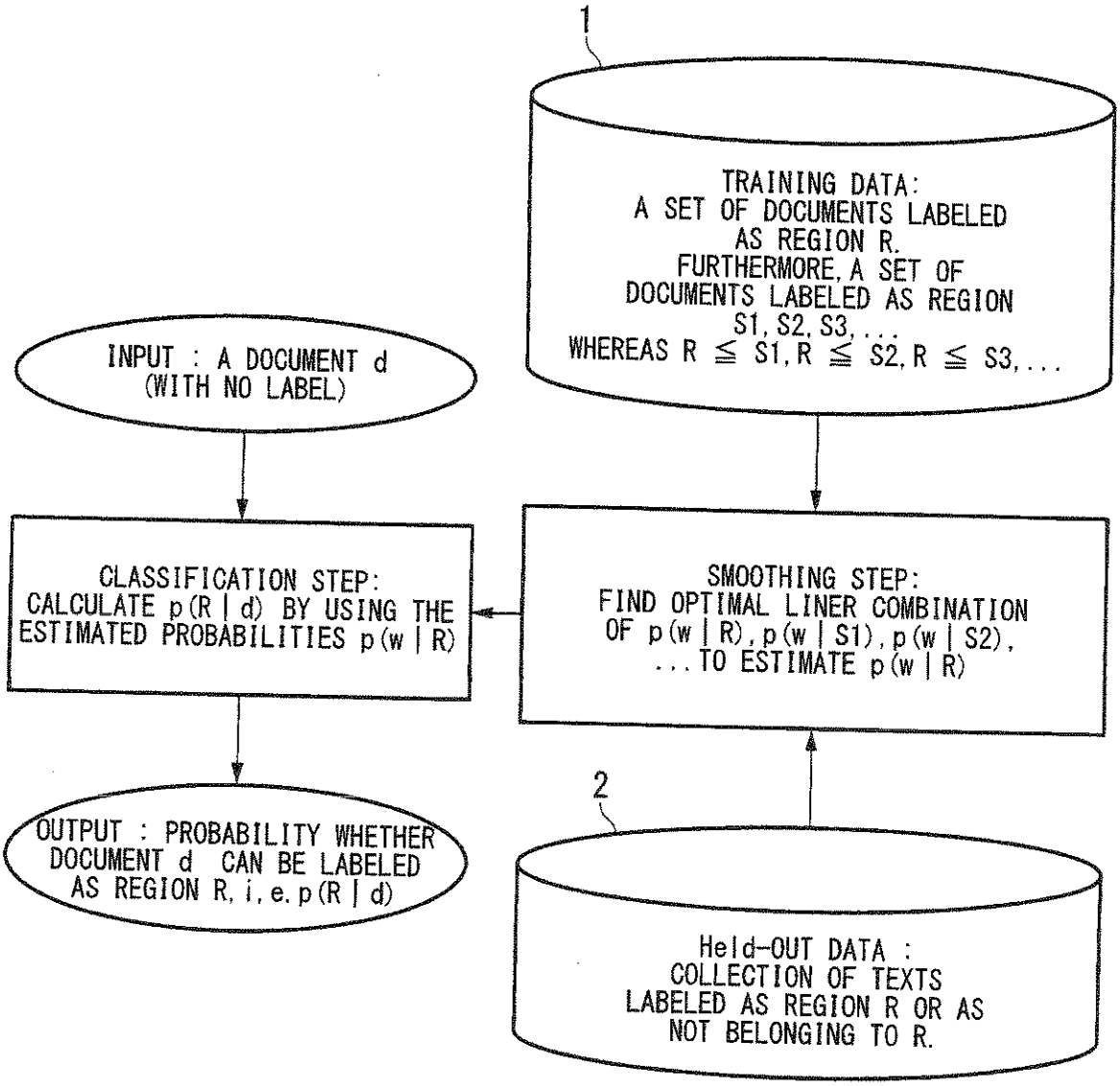

1
TRAINING DATA:
A SET OF DOCUMENTS LABELED AS REGION R.
FURTHERMORE, A SET OF DOCUMENTS LABELED AS REGION S1, S2, S3, ...
WHEREAS R ≦ S1, R ≦ S2, R ≦ S3, ...
INPUT : A DOCUMENT d (WITH NO LABEL)
CLASSIFICATION STEP:
CALCULATE p(R | d) BY USING THE ESTIMATED PROBABILITIES p(w | R)
SMOOTHING STEP:
FIND OPTIMAL LINER COMBINATION OF p(w | R), p(w | S1), p(w | S2), ... TO ESTIMATE p(w | R)
OUTPUT : PROBABILITY WHETHER DOCUMENT d CAN BE LABELED AS REGION R, i.e. p(R | d)
2
Held-OUT DATA :
COLLECTION OF TEXTS LABELED AS REGION R OR AS NOT BELONGING TO R.

FIG. 2

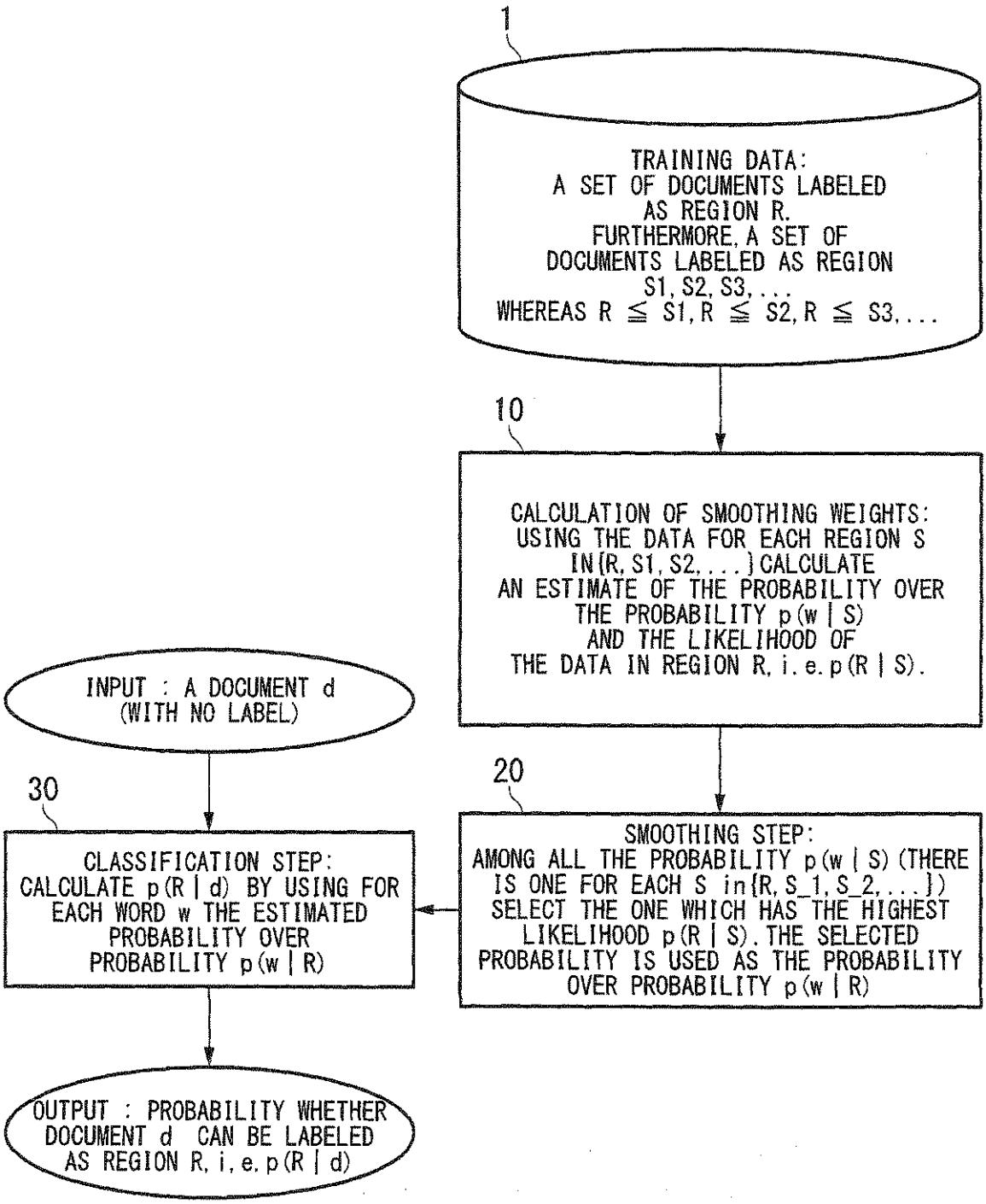

1
TRAINING DATA:
A SET OF DOCUMENTS LABELED AS REGION R.
FURTHERMORE, A SET OF DOCUMENTS LABELED AS REGION S1, S2, S3, ...
WHEREAS R ≦ S1, R ≦ S2, R ≦ S3, ...
10
CALCULATION OF SMOOTHING WEIGHTS:
USING THE DATA FOR EACH REGION S IN{R, S1, S2, ...} CALCULATE AN ESTIMATE OF THE PROBABILITY OVER THE PROBABILITY p(w | S) AND THE LIKELIHOOD OF THE DATA IN REGION R, i.e. p(R | S).
20
SMOOTHING STEP:
AMONG ALL THE PROBABILITY p(w | S) (THERE IS ONE FOR EACH S in{R, S_1, S_2, ...}) SELECT THE ONE WHICH HAS THE HIGHEST LIKELIHOOD p(R | S). THE SELECTED PROBABILITY IS USED AS THE PROBABILITY OVER PROBABILITY p(w | R)
INPUT : A DOCUMENT d (WITH NO LABEL)
30
CLASSIFICATION STEP:
CALCULATE p(R | d) BY USING FOR EACH WORD w THE ESTIMATED PROBABILITY OVER PROBABILITY p(w | R)
OUTPUT : PROBABILITY WHETHER DOCUMENT d CAN BE LABELED AS REGION R, i.e. p(R | d)

FIG. 3

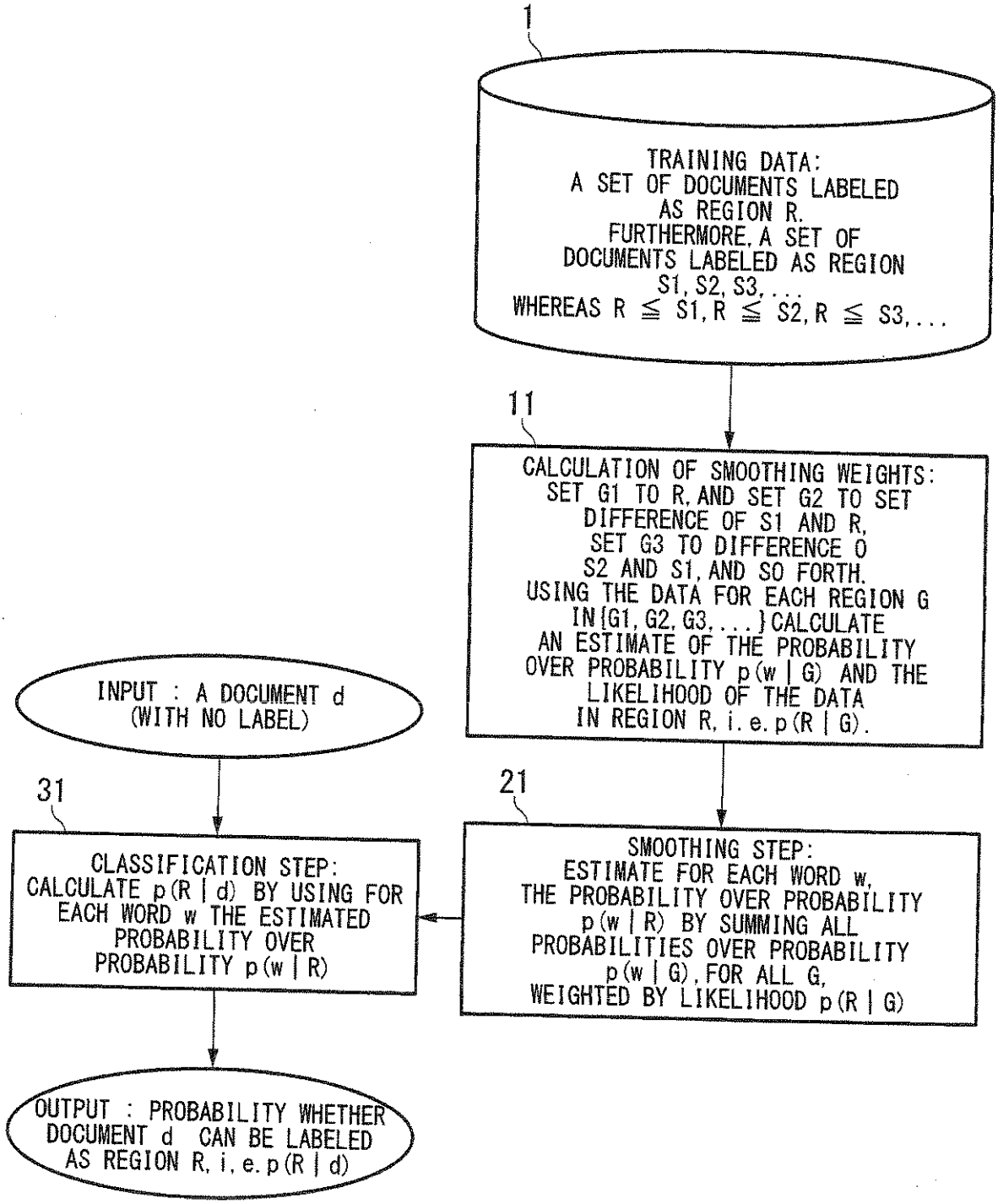

1
TRAINING DATA:
A SET OF DOCUMENTS LABELED AS REGION R.
FURTHERMORE, A SET OF DOCUMENTS LABELED AS REGION S1, S2, S3, ...
WHEREAS R ⊆ S1, R ⊆ S2, R ⊆ S3, ...
11
CALCULATION OF SMOOTHING WEIGHTS:
SET G1 TO R, AND SET G2 TO SET DIFFERENCE OF S1 AND R,
SET G3 TO DIFFERENCE OF S2 AND S1, AND SO FORTH.
USING THE DATA FOR EACH REGION G IN {G1, G2, G3, ...} CALCULATE AN ESTIMATE OF THE PROBABILITY OVER PROBABILITY p(w | G) AND THE LIKELIHOOD OF THE DATA IN REGION R, i.e. p(R | G).
21
SMOOTHING STEP:
ESTIMATE FOR EACH WORD w, THE PROBABILITY OVER PROBABILITY p(w | R) BY SUMMING ALL PROBABILITIES OVER PROBABILITY p(w | G), FOR ALL G, WEIGHTED BY LIKELIHOOD p(R | G)
31
CLASSIFICATION STEP:
CALCULATE p(R | d) BY USING FOR EACH WORD w THE ESTIMATED PROBABILITY OVER PROBABILITY p(w | R)
INPUT : A DOCUMENT d (WITH NO LABEL)
OUTPUT : PROBABILITY WHETHER DOCUMENT d CAN BE LABELED AS REGION R, i.e. p(R | d)

FIG. 4

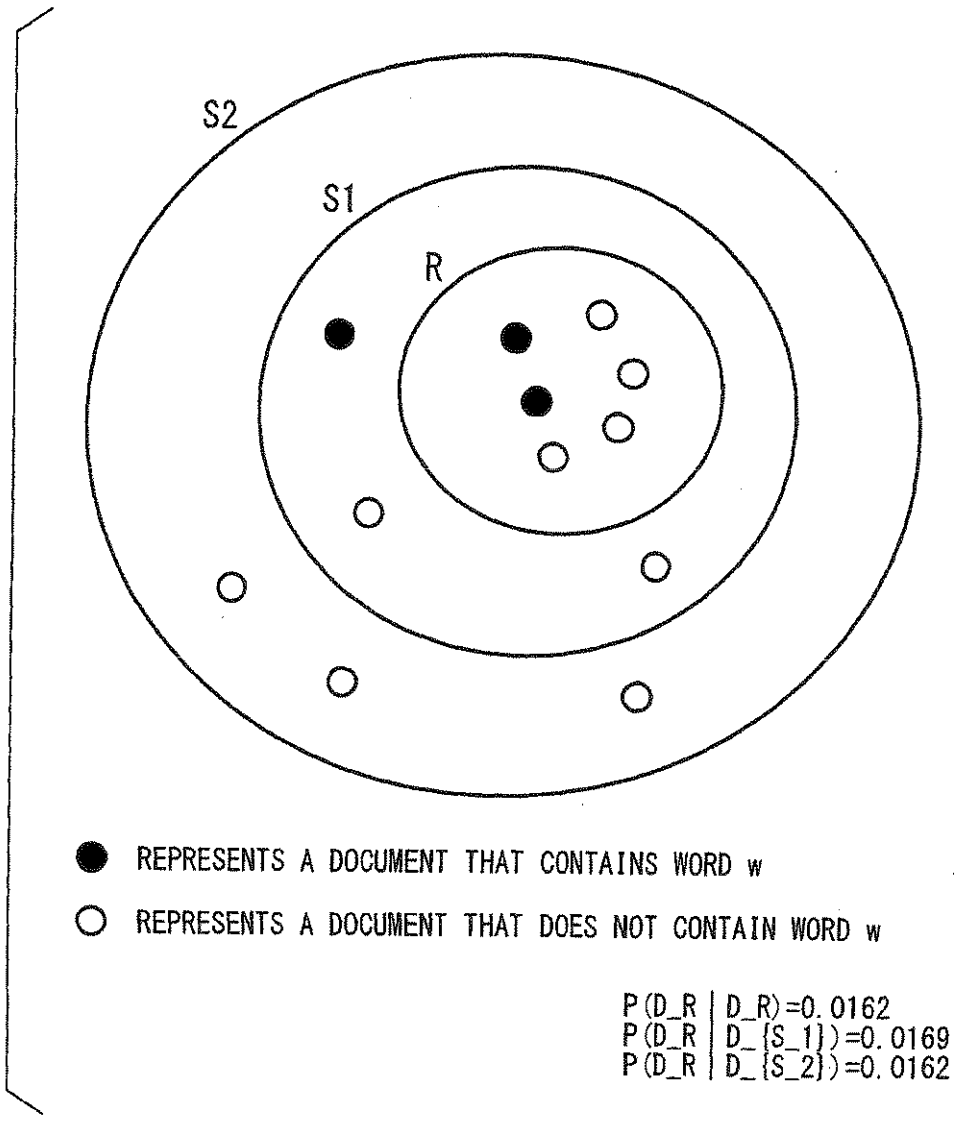
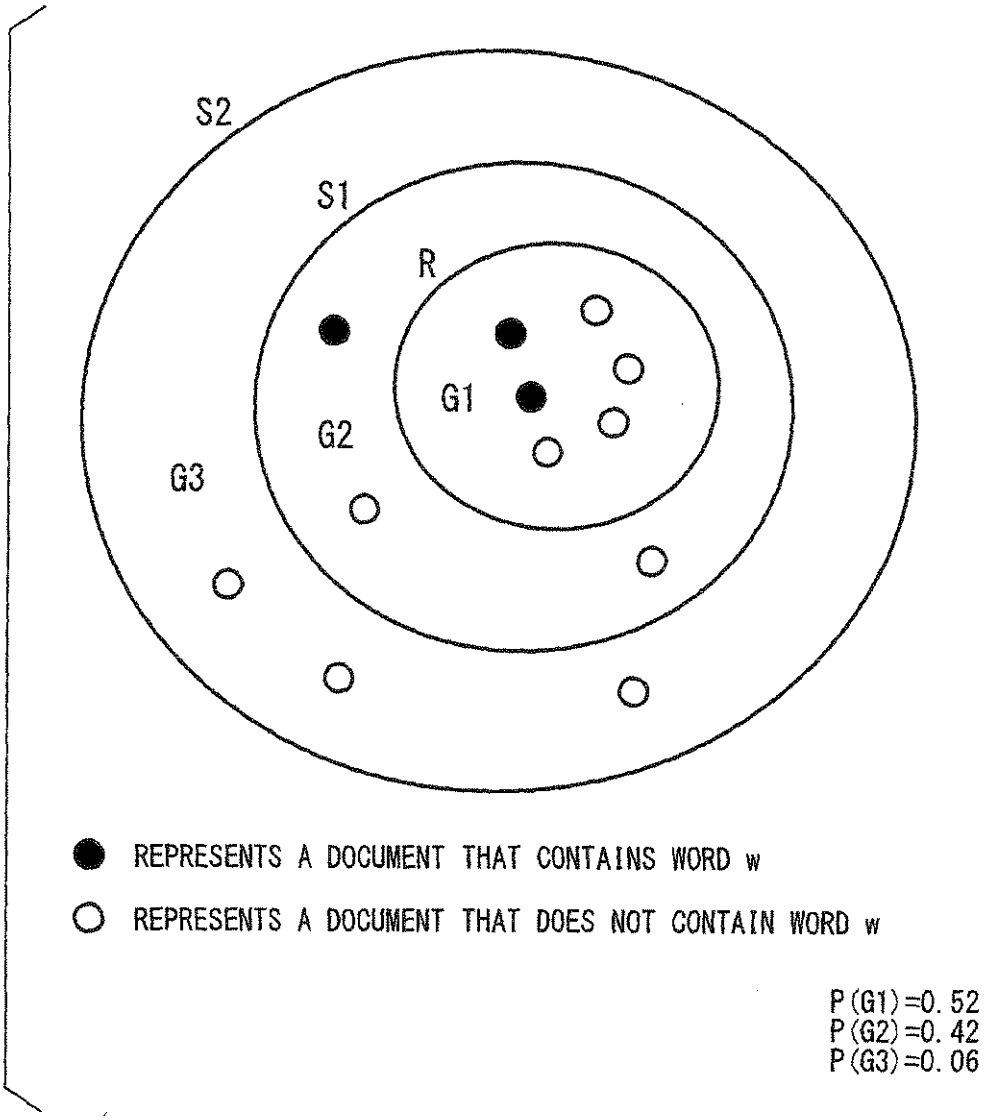

S2
S1
R
● REPRESENTS A DOCUMENT THAT CONTAINS WORD w
○ REPRESENTS A DOCUMENT THAT DOES NOT CONTAIN WORD w
P(D_R | D_R)=0.0162
P(D_R | D_{S_1})=0.0169
P(D_R | D_{S_2})=0.0162

FIG. 5


S2
S1
R
G1
G2
G3
● REPRESENTS A DOCUMENT THAT CONTAINS WORD w
○ REPRESENTS A DOCUMENT THAT DOES NOT CONTAIN WORD w
P(G1)=0.52
P(G2)=0.42
P(G3)=0.06