Performance Evaluation and Visualization of Association Rules using Receiver Operating Characteristic Graph

Minoru Kawahara
Data Processing Center
Kyoto University
Kyoto 6068501, Japan
kawahara@kudpc.kyoto-u.ac.jp

Hiroyuki Kawano
Department of Systems Science
Kyoto University
Kyoto 6068501, Japan
kawano@i.kyoto-u.ac.jp

Abstract

We have been developing the web search engine, "mondou", *using weighted association rules. It is very helpful for search users to provide associative keywords which are derived by text mining algorithm. Moreover, based on the experimental results of our web search system, we try to implement these mining functions on web-based intelligent database navigation system using the source program files of the commercial text database "OpenText". Especially, in order to derive appropriate keywords with small computing cost, we carefully focus on how to determine system parameters, such as $Minsup$ and $Minconf$ threshold values. In this paper, we use the techniques of ROC graph to evaluate the performance and characteristics of derived rules. We propose the ROC analytical model of our search system, and evaluate the performance of weighted association rules by ROC convex hull method. Moreover, we also propose a method which visualizes association rules using ROC graph to provide the relationship between a query and derived association rules. By using the INSPEC database, we try to specify the optimal threshold values to derive effective rules from the typical bibliographic data.*

1. Introduction

In the research fields of data mining[2, 3] and text mining [10], various algorithms have been proposed to discover interesting patterns, rules, trends and representations in various databases. Many algorithms only can derive quite simple patterns or rules as knowledge, it is very hard to derive meaningful rules or knowledge in the viewpoints of human experts. However, even these simple patterns or rules may be helpful for beginners, who don't have much background or domain knowledge of the interesting topics.

Therefore, we focus on the basic algorithm of association rules [11], we have been developing intelligent web search engine "Mondou" (http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/). Our Mondou style navigation system, which provides associative keywords to users, provides very helpful functions to beginners[7, 8]. Moreover, extending the algorithm of Mondou web search engine, we have been implementing the web-based bibliographic navigation interface[5, 6], using the source program files of the commercial text database "OpenText".

In Figure 1 and 2, we present the web interface of our Mondou style bibliographic navigator, which provides associative keywords derived in bibliographic databases.

However, it is rather difficult for the administrator to determine system parameters, such as minimum support threshold $Minsup$ and minimum confidence threshold $Minconf$. Hence, by our proposed algorithms[5, 6], we try to adjust the $Minsup$ values dynamically according to the characteristics of stored documents. Our basic idea of the implementing algorithm is based on the comparison of different attribute values among several categorized attributes of bibliographic databases.

Shortly speaking, if no keyword is derived using a given $Minsup$ value, then we increase $Minsup$ value step by step till the upper limitation given by the administrator. We also use the characteristics of intersection for keyword sets derived from different attributes, such as `Title` and `Author` attributes. Furthermore, if the number of derived keywords is more than the maximum threshold $Maxkey$, we have to decrease $Minsup$ value in order to derive the less number of keywords.

Then, in this paper, we propose a method which specifies the optimal thresholds based on the ROC (Receiver Operating Characteristic) analysis[1, 9] and evaluate the performance of our proposed method based on the experimental results from our Mondou style bibliographic navigation. And we also propose a method which visualizes association rules using the ROC to provide the relationship between a

0-7695-0496-5/00 $10.00 © 2000 IEEE

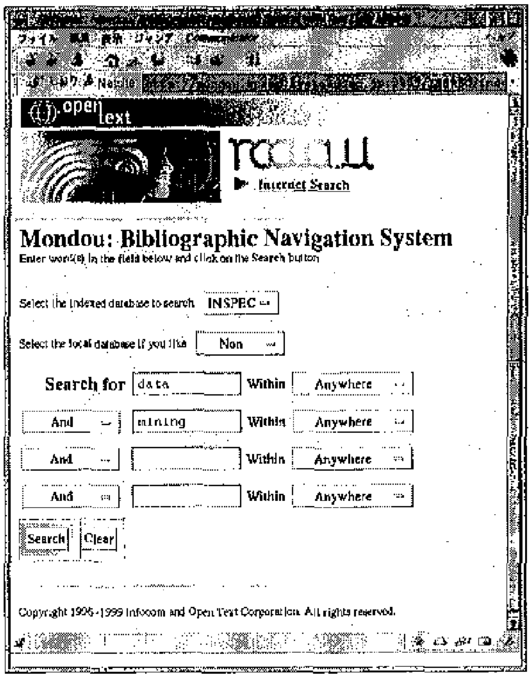

Figure 1. The query window of Mondou **style system.**

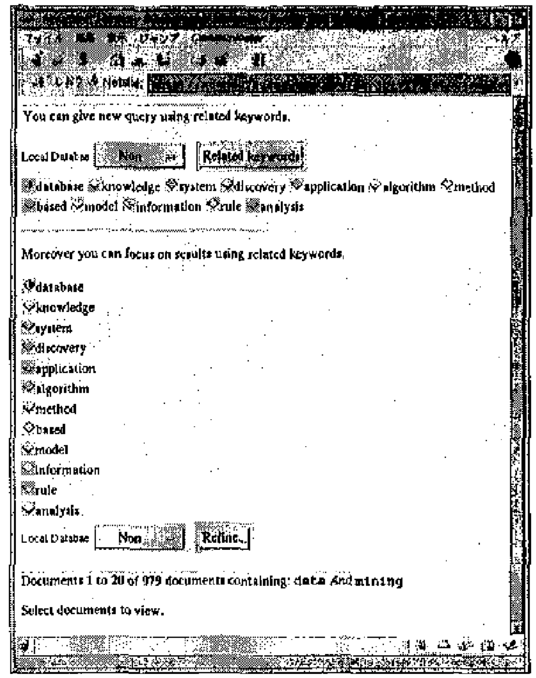

Figure 2. The results of Mondou **style system.**

query and derived association rules.

In Section 3 we define several parameters for the ROC analysis and propose a model in order to evaluate the performance of our bibliographic navigation system. In Section 4 we evaluate the performance of our algorithm based on the experimental results of our INSPEC navigation system. In Section 5 we describe about visualization of derived rules using ROC. Finally, we make concluding remarks and discuss the future works in Section 6.

2. ROC Analysis

ROC graphs have been used in the signal detection theory to depict tradeoffs between the hit rate and the false alarm rate. ROC graphs illustrate the behavior of a classifier without regard to class distribution or error cost, and so they decouple classification performance from these factors. Moreover, the ROC convex hull method is a way to compare multiple classifiers on an ROC graph, and it specify the optimal classifier which supplies the highest performance[9, 1].

2.1. ROC Graph

It is assumed that a instance can be classified into two instance classes: the positive instance class P or the negative instance class N, and positive y (yes) or negative n (no) are assigned by a classifier.

We also assume that $p(c \mid I)$ is the posterior probability that instance I is positive c. Then the true positive rate TP of a classifier is:

$$TP \quad = \quad p(y \mid P) \simeq \frac{\text{positive correctly classified}}{\text{total positives}}. \quad (1)$$

On the other hand, the false positive rate FP of a classifier is:

$$FP \quad = \quad p(y \mid N) \simeq \frac{\text{negative incorrectly classified}}{\text{total negatives}} \quad (2)$$

When we focus on the concepts of ROC curves in an ROC graph, FP is plotted on the X axis and TP is plotted on the Y axis on a graph for several instances, we can draw monotone curves which are shown in Figure 3.

Moreover, TP is higher and the point is located in the upper area of the ROC graph, it represents that a instance is classified correctly by the classifier. The right point (FP is higher) represents that a instance is classified incorrectly by the classifier. Therefore the ROC curve near the area of higher TP and lower FP, that is the most-northwest line, must be better. Therefore, we can conclude that the curve A is better than curve D, because it dominates in all points.

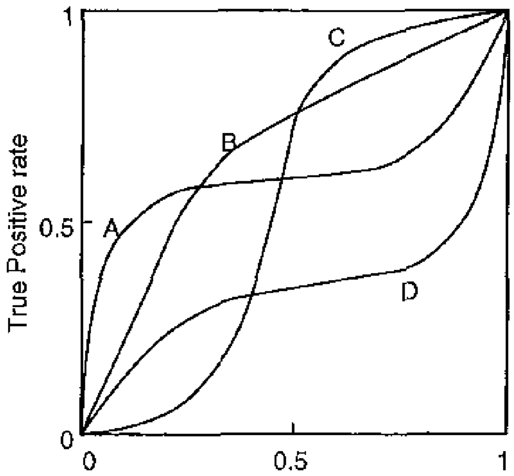

Figure 3. An ROC graph of four classifiers

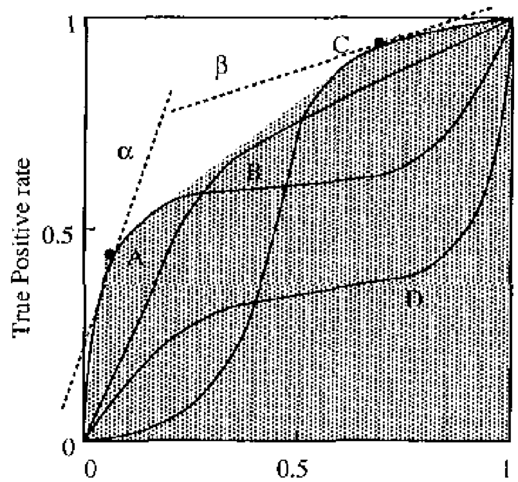

Figure 4. Lines α and β show the optimal classifier under different sets of conditions.

2.2. ROC Convex Hull Method

In this subsection, we try to consider the evaluation of the different classifiers, since ROC graphs illustrate the performance curves which depend on only the accuracy of classifiers.

Let's assume that $c(classification, class)$ is a error cost function, and that $c(n, P)$ is the cost of a false negative error and $c(y, N)$ is the cost of a false positive error. It is also assumed that $p(P)$ is the prior probability of a positive instance, so the prior probability of a negative instance is $p(N) = 1 - p(P)$.

Then the expected cost of a classification by the classifier represented by a point (FP, TP) on the ROC graph is:

$$p(P) \cdot (1 - TP) \cdot c(n, P) + p(N) \cdot FP \cdot c(y, N). \tag{3}$$

If two points, (FP_1, TP_1) and (FP_2, TP_2), have the same performance, then we have the following equation:

$$\frac{TP_2 - TP_1}{FP_2 - FP_1} = \frac{p(N) \cdot c(y, N)}{p(P) \cdot c(n, P)}. \tag{4}$$

This equation gives the slope of an iso-performance line through two points, (FP_1, TP_1) and (FP_2, TP_2), which have the same performance. Hence, the slope of an iso-performance line is specified by $p(N)/p(P)$ and the ratio of error $c(y, N)/c(n, P)$.

For example, the situation having $p(N)/p(P) = 3$, if false negative and false positive errors have equal cost, then the slope of iso-performance lines is 3. And if a false negative value takes 10 times expensive of a false positive value, then the slope of iso-performance lines is $3/10$. Then the best classifier must be located on the most-northwest of iso-performance line. Therefore, a classifier is potentially optimal if it lies on the north-west boundary of the convex hull drawn as the border between the shaded and unshaded areas in Figure 4.

For example, in Figure 4, α is the best line with slope 3 that touches the convex hull, and A is the best classifier because A constructs the line. On the other hand, β is the best line with slope $1/3$ that touches the convex hull constructed by C, and we can conclude that C is the best classifier in that case. Consequently, although the ROC curve drawn by B have partly better performance than A and C, the better performance curve is dominated by the convex hull constructed by only A and C.

3. ROC Analytical Model of Bibliographic Navigation System

In this section, we apply the ROC analysis to **Mondou** style bibliographic navigation system. Let's assume that $\bigcup$ is the set operator of union, $\bigcap$ is the set operator of intersection, and $||$ is the set operator to count the number of items. Moreover, we will define following parameters:

Definition

- $\mathbf{G}$: A set of keywords including in a query
- n: The number of keywords in $\mathbf{G}$
- k_i: The i'th keyword in $\mathbf{G}$ $(1 \leq i \leq n)$

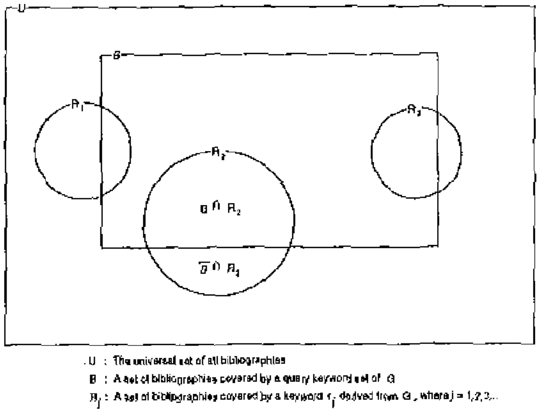

Figure 5. Status of bibliographies covered by keywords in a bibliographic database.

$\mathbf{K}_i$: The set of bibliographies covered by k_i

$\mathbf{B}$: The set of bibliographies covered by $\mathbf{G}$

m: The number of keywords derived from $\mathbf{G}$

r_j: The j'th keyword derived from $\mathbf{G}$ $(1 \leq j \leq m)$

$\mathbf{R}_j$: The set of bibliographies covered by r_j

Figure 5 shows a status of coverage by $\mathbf{B}$ and $\mathbf{R}_j$ in the universal set U, which is all bibliographies in the system. The bibliography set covered by all keywords in $\mathbf{G}$ is:

$$\mathbf{B} = \bigcap_{i=1}^{n} \mathbf{K}_i. \tag{5}$$

In a retrieved set, the positive instance is those $\mathbf{B}$ that decreases the number of bibliographies and the negative instance is those $\overline{\mathbf{B}}$ that increases the number of them. Thus the true positive instance is that $\mathbf{B} \cap \bigcup_{j=1}^{m} \mathbf{R}_j$, and the false positive instance is that $\overline{\mathbf{B}} \cap \bigcup_{j=1}^{m} \mathbf{R}_j$. Then, TP is represented by:

$$TP = \frac{|\mathbf{B} \cap \bigcup_{j=1}^{m} \mathbf{R}_j|}{|\mathbf{B}|}, \tag{6}$$

and FP is represented by :

$$FP = \frac{|\overline{\mathbf{B}} \cap \bigcup_{j=1}^{m} \mathbf{R}_j|}{|\overline{\mathbf{B}}|}. \tag{7}$$

For instance, in Figure 5, the true positive instance is $\mathbf{B} \cap \mathbf{R}_2$ and the false positive instance is $\overline{\mathbf{B}} \cap \mathbf{R}_2$.

Then, using the definitions of TP and FP, we illustrate ROC graphs by plots of points (FP, TP) using different $Minsup$ values as classifiers. By applying the ROC convex hull method to the ROC graphs, we can choose the best classifier based on the value of $Minsup$.

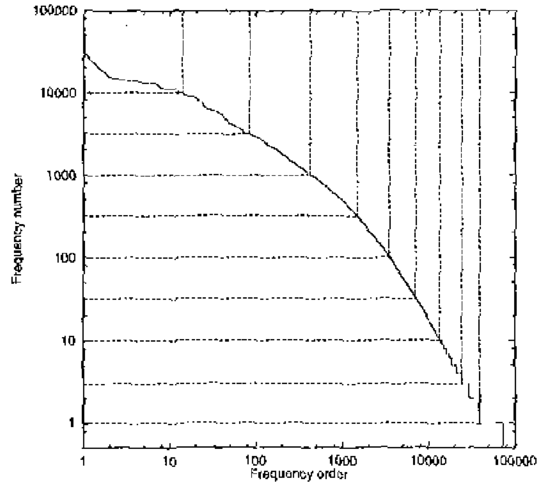

Figure 6. The frequency of the commonly used words and the categories separated by lines.

4. Performance Evaluation of Navigation System for INSPEC Database

Our Mondou system is able to derive association rules in full INSPEC database. Our implementing bibliographic navigation system can handle 3,012,864 items, which have been published by INSPEC from January 1987 to December 1997. In this section, we focus on 330,562 titles in 1997 in order to evaluate the performance of our proposed algorithm. Hence, the number of all bibliographies in the retrieved set, | U |, is 330,562.

In Figure 6, we present the the frequency of commonly used keywords in `Title` attribute. Of course, we pay an attention to remove meaningless keywords, such as "of", "the" and "and", so it is required to avoid the derivation of those stop keywords.

Moreover, based on the log scale of frequency of keywords, we equally divide into several classes in order to draw curves on an ROC graph, which is shown in Figure 6. We plot averages of (FP, TP) derived from specific keywords, considering the categories in an ROC graph. Categories in Table 1 are equivalent to each area in Figure 6. Then we make random samples of keywords having more than 1 % of keywords in each category according to the "sampling rate", which is shown in Table 1.

In order to avoid the unstability of an interaction of $Minconf$ and $Minsup$ in deriving associative keywords, we try to fix $Minconf = 0.01$, and we change $Minsup$ gradually. The average numbers of keywords derived from the sampled keywords are shown in Figure 7 for each category. It is clear that the lower $Minsup$ becomes, the more

Table 1. Categories of retrieved keywords.

category	frequency			number of keywords	sampling rate
1	10001	~		13	13
2	3163	~	10000	70	20
3	1001	~	3162	341	20
4	317	~	1000	1048	20
5	101	~	316	1974	20
6	33	~	100	3562	36
7	11	~	32	6302	64
8	4	~	10	10722	108
9	2	~	3	14540	146
10	1			34738	348

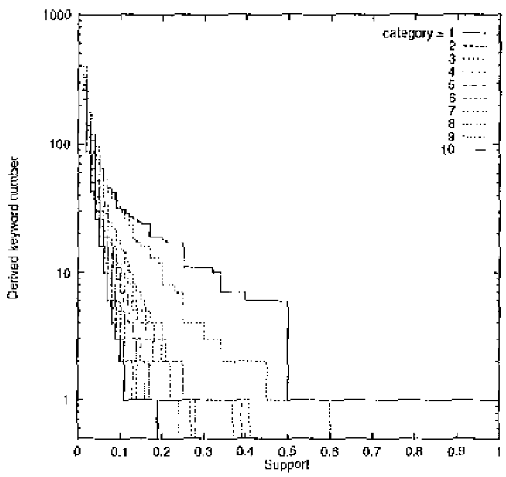

Figure 7. Average number of derived keywords for each category.

rapidly the number of derived keywords becomes. It is also shown that higher frequency of keywords are, the lower *support* of the keywords is, so the numbers of derived keywords are very different for each category if $Minsup$ is fixed.

In order to draw ROC curves by taking different $Minsup$ values as classifiers, we have to determine the values of *support*. However, we have to pay an attention that there is no derived keyword for some categories, when we take small values of *support* which is less than 0.2. Then we try to evaluate the system based on the following $Minsup$:

$$Minsup = \{0.02, 0.04, 0.06, 0.08, 0.1, 0.15, 0.2, 0.25, 0.30, 0.4, 0.5, 0.6\}.$$

By using the above $Minsup$, we show the numbers of derived keywords in Figure 8, and that more number of keywords tends to be derived from keywords with lower

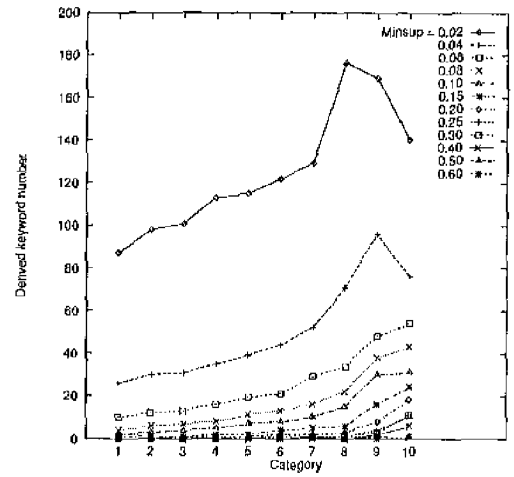

Figure 8. Average number of derived keywords by $Minsup$.

frequency. Thus, it is able to conjecture that TP will be higher derived from a lower frequency keywords, because derived keywords are increased to cover more the bibliographic items. Moreover Figure 10 and Figure 11 show that TP is rather than bigger than FP values according to the $Minsup$–FP and $Minsup$–TP curves. However, it may be impossible for users to select keywords effectively, if many associative keywords are derived. So it is required to control the value of $Minsup$ according to the frequency of the keywords appeared in the query.

Figure 9 shows the ROC graph plotted by the averages of (FP, TP) for each category and each $Minsup$ with the boundary of the convex hull. Table 2 shows the best classifiers for each slope of iso-performance line, that is the best $Minsup$, in Figure 9. In Table 2, "AllPos" represents to derive all associative keywords and "AllNeg" represents to derive no associative keyword.

As shown in Figure 9, the ROC convex hull was pulled by a part of plots near the upper area and warped. It is also found that there is a jumping point where $Minsup$ takes the value from 0.02 to 0.08 in Table 2 . It is also clear that lines in 8, 9, 10 categories converge, which are shown in Figure 10 and Figure 11.

This is the reason that such keywords in those categories only appear less than 10 times in the target INSPEC database, and that it is impossible to derived effective keywords from so few tuples. Therefore, in Figure 12, we present the ROC graph without such categories and show Table 3 having the optimal $Minsup$ values.

By the way, $p(N)/p(P)$ of a keyword can be represented

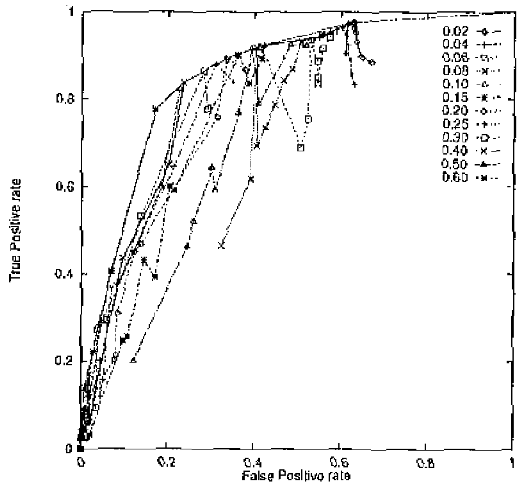

Figure 9. The ROC graph of $Minsup$ with the iso-performance line.

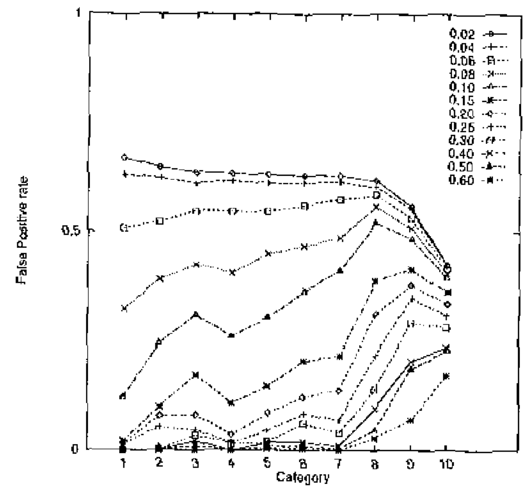

Figure 10. Values of FP for each category.

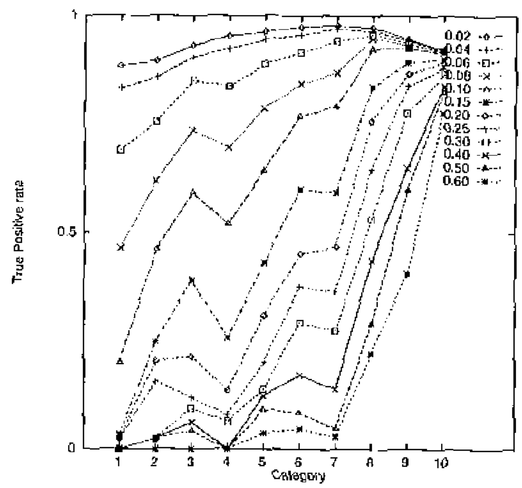

Figure 11. Values of TP for each category.

by the following equation:

$$\frac{p(N)}{p(P)} = \frac{|\mathbf{U}| - |\mathbf{B}|}{|\mathbf{B}|}, \tag{8}$$

where $\mathbf{B}$ represents a set of bibliographic items covered by the keyword. $|\mathbf{B}|$ is given as the hit counts retrieved by the keyword, and $|\mathbf{U}|$ is 330, 562.

The ratio of error cost $c(y,N)/c(n,P)$ may be given by the system administrator. In this experience, we determine values of the cost not to derive any keywords for the meaningless categories from 8 to 10. From Table 3, when we have the value of slope (4) is not less than 227.06, it makes classifier to be AllNeg. Hence

$$\frac{|\mathbf{U}| - |\mathbf{B}|}{|\mathbf{B}|} \cdot \frac{c(y,N)}{c(n,P)} = 227.06$$

leads the following equation:

$$R_{error} = \frac{c(n,P)}{c(y,N)} = \frac{330562 - |\mathbf{B}|}{227.06 \times |\mathbf{B}|},$$

where R_{error} is the error cost ratio of the false positive and true negative so that the more expensive R_{error} is, the fewer retrieval omission is.

If R_{error} is substituted $|\mathbf{B}| = \{10, 1\}$ for, then R_{error} becomes $\{145, 1455\}$. From these R_{error} values, $Minsup$ are specified as shown in Table 4 and Table 5. Based on such values, the number of derived keywords (keyword number), the ratio of no derived keyword and the sampled keywords (NDK) and the average value of $Minsup$ are shown in Table 6.

The result by using our basic algorithm[5, 6] is also shown in Table 6, where our basic algorithm reduces the derived keywords, if keywords are derived more than the maximum threshold $Maxkey$ as described in Section 1.

Comparing the number of keywords in Table 6, it is found that the more frequent keywords in a query are, the fewer derived keywords are by the basic proposed algorithm, but the result is opposite to it by the ROC algorithm. We have the following reason:

By the mining association algorithm, if association rules are derived from a highly frequent keywords, then there are many retrieved bibliographies and the *support* is fewer relatively. By the basic algorithm, $Minsup$ is fixed so that only keywords which are frequent cooccurrent of the query keywords can exceed the threshold. As a result, a few keywords are derived from such frequent keywords.

By using the proposed ROC algorithm, if the frequency

Table 6. The derived result by our basic algorithm and ROC algorithms: NDK stands for "no derived keyword percentage"

Category	Basic Algorithm $Support = 0.08$			ROC Algorithm $R_{error} = 145$			ROC Algorithm $R_{error} = 1455$		
	keyword number	NDK [%]	$Minsup$ average	keyword number	NDK [%]	$Minsup$ average	keyword number	NDK [%]	$Minsup$ average
1	5	0	0.08	89	0	0.02	265	0	AllPos
2	6	0	0.08	58	0	0.03	227	0	AllPos
3	8	0	0.08	4	35	0.17	103	0	0.02
4	9	0	0.08	2	80	0.34	82	0	0.03
5	10	0	0.08	2	95	0.50	5	15	0.16
6	12	0	0.09	2	94	0.60	2	56	0.34
7	12	0	0.10	1	94	0.60	1	88	0.49
8	8	4	0.18	0	100	AllNeg	2	64	0.60
9	8	13	0.31	0	100	AllNeg	3	36	0.60
10	10	3	0.27	0	100	AllNeg	0	100	AllNeg

Table 2. The optimum $Minsup$ derived from the ROC convex hull.

Slope	Classifier ($Minsup$)
0.0000 ~ 0.0638	AllPos
0.0638 ~ 0.2597	0.02
0.2597 ~ 0.2746	0.08
0.2746 ~ 0.4126	0.10
0.4126 ~ 0.4535	0.20
0.4535 ~ 0.5478	0.25
0.5478 ~ 0.5751	0.30
0.5751 ~ 0.9644	0.40
0.9644 ~ 3.6601	0.60
3.6601 ~ 4.4116	0.60
4.4116 ~ 12.906	0.40
12.906 ~ 227.06	0.60
227.06 ~	AllNeg

of query keywords is lower, then the value of $p(N)/p(P)$ becomes less so that the threshold is lower such value of $Minsup$ as shown in Table 4 and more keywords can exceed it to be derived.

It is also that the basic algorithm derives highly frequent keywords from very popular keywords, but the ROC algorithm can also derives low frequent keywords.

For example, by the basic algorithm with $Minsup = 0.08$, the keywords derived from "performance", which belongs to category 2 and appears 5,536 times, and to which the following 8 keywords are highly associative:

> system(1), high(1), simulation(2), model(1), control(1), evaluation(2), analysis(1), network(1),

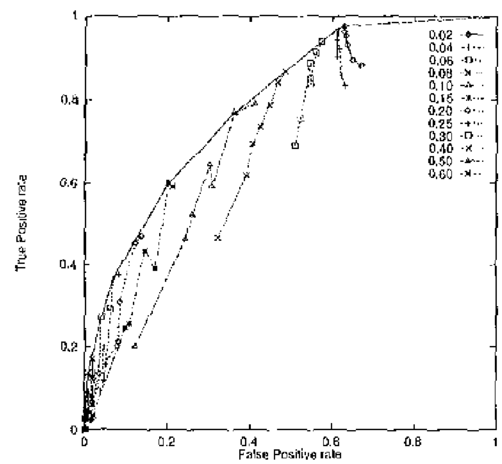

Figure 12. The ROC graph of $Minsup$ with the iso-performance line without categories from 8 to 10.

where the number in braces represents the category to which the keywords belong, and all of derived keywords belong to category 1 or 2, which hold high frequency of keywords.

By the ROC algorithm, $Minsup$ becomes 0.02 to derive additional keywords except of the above mentioned keywords and 71 keywords are derived totally:

> *time(2), management(3), design(2), computer(3),* assessment(3), machine(3), data(2), process(2), method(1), based(1), algorithm(2), parallel(2), effect(1), processing(2), optical(2), broadband(4)

There are 10 keywords in category 1, 24 keywords in cate-

Table 3. The optimum $Minsup$ derived from the ROC convex hull without categories from 8 to 10.

Slope	Classifier ($Minsup$)
$0.0000 \sim 0.0638$	AllPos
$0.0638 \sim 0.4791$	0.02
$0.4791 \sim 0.7195$	0.04
$0.7195 \sim 0.8103$	0.06
$0.8103 \sim 1.0589$	0.10
$1.0589 \sim 1.7351$	0.15
$1.7351 \sim 3.2032$	0.25
$3.2032 \sim 4.3497$	0.30
$4.3497 \sim 12.906$	0.40
$12.906 \sim 227.06$	0.60
$227.06 \sim$	AllNeg

Table 4. $Minsup$ **at** $R_{error} = 145$.

Category	$p(N)/p(P) \cdot 1/R_{error}$	Optimal $Minsup$
1	$0.0000 \sim 0.2211$	AllPos $\sim$ 0.02
2	$0.2211 \sim 0.7139$	$0.02 \sim 0.04$
3	$0.7141 \sim 2.2706$	$0.04 \sim 0.25$
4	$2.2728 \sim 7.1847$	$0.25 \sim 0.40$
5	$7.2075 \sim 22.565$	$0.40 \sim 0.60$
6	$22.790 \sim 69.076$	0.60
7	$71.235 \sim 207.24$	0.60
8	$227.97 \sim 569.93$	AllNeg
9	$759.91 \sim 1139.9$	AllNeg
10	2279.7	AllNeg

gory 2, 29 keywords in category 3, and 8 keywords in category 4. It is clear that there are various keywords which have relatively low frequency in the category 3 and 4.

Next, we show the table 6 which presents the average distance between points (FP, TP) and the point $(1, 0)$. On ROC graphs, when the distance is longer from the point $(1, 0)$, the performance becomes higher. We conclude that the ROC algorithm shows higher performance than the basic algorithm on any R_{error} values. Therefore the ROC algorithm can derive low frequent keywords even though the basic algorithm derives only highly frequent keywords which are little significant as knowledge[5].

Using the ROC algorithm, we can derive more helpful keywords which are guaranteed by the ROC analysis. And even if such algorithms that suppress highly frequent keywords are added to the mining algorithm, we can discover more effective keywords which are tightly associative to input queries.

Moreover, when we focus on the gap of 0% and non 0% in Table 6, we have to pay an attention that the numbers of derived keywords is quite different. It seems that the frequencies of keywords, which effect on $p(N)/p(P)$, have a weak boundary. Thus, when we specifies R_{error} dynamically according to the frequencies of query keywords, it is very effective to apply our proposed method to determine the threshold values even in any situations.

Table 5. $Minsup$ **at** $R_{error} = 1455$.

Category	$p(N)/p(P) \cdot 1/R_{error}$	Optimal $Minsup$
1	$0.0000 \sim 0.0220$	AllPos
2	$0.0220 \sim 0.0711$	AllPos $\sim$ 0.02
3	$0.0712 \sim 0.2263$	0.02
4	$0.2265 \sim 0.7160$	$0.02 \sim 0.04$
5	$0.7183 \sim 2.2487$	$0.04 \sim 0.25$
6	$2.2712 \sim 6.8839$	$0.25 \sim 0.40$
7	$7.0990 \sim 20.653$	$0.40 \sim 0.60$
8	$22.718 \sim 56.797$	0.60
9	$75.729 \sim 113.59$	0.60
10	227.19	AllNeg

5. Visualization of Association Rules

Our current navigation system enumerates keywords derived from a query by the method described in Section 2 as a keyword list on a Web browser. Thus it is required for query users to select appropriate keywords in the list by their themselves and improve input queries. But it is difficult for them how to select appropriate keywords in the list which only enumerates keywords and doesn't show the relationship between the query and the derived keywords.

So we propose a method to plot the derived keywords on a ROC graph as a ROC map according to each value of FP and TP and visualize the relationship between query and derived keywords. On the ROC graph a keyword nearer to the most north line $TP = 1$ covers the same space as the query and a keyword nearer to the most west line $FP = 1$ covers the different space from the query in the retrieval. So it is thought that TP shows homogeneousness and FP shows heterogeneousness as shown in Figure 13. Therefore a keyword near to the most north west point $(0, 1)$ has low heterogeneousness and high homogeneousness, so it is closely associative to the query. They may be words in an idiom. On the other hand a keyword near to the most south east point $(1, 0)$ has high heterogeneousness and low homogeneousness, so it is unassociative to the query. They may be used in different field in the world. For example, Figure 14 shows a ROC map derived from "replanning" and Figure 15 shows one derived from "performance" where $R_{error} = 145$. "replanning" appears only 6 times in Title and belongs to the category 8. and "performance" appears 5,536 times in Title and belongs to the category 2.

Moreover, ROC map can show another rules than the

Table 7. The average distances from the point $(1, 0)$ on the ROC graph.

Category	Basic Algorithm ($Support = 0.08$)	ROC Algorithm ($R_{error} = 145$)		ROC Algorithm ($R_{error} = 1455$)	
	distance	distance	remainder	distance	remainder
1	0.8211	0.9477	0.1266	0.9510	0.1299
2	0.8940	0.9725	0.0785	0.9791	0.0851
3	0.9119	0.9008	-0.0111	1.0056	0.0937
4	0.9322	0.9857	0.0535	1.0156	0.0834
5	0.9926	0.9976	0.0050	0.9541	-0.0385
6	0.9929	0.9968	0.0039	0.9852	-0.0077
7	1.0262	1.0001	-0.0261	1.0013	-0.0249
8	1.0159	1.0000	-0.0159	0.9940	-0.0219
9	1.0351	1.0000	-0.0351	1.0229	-0.0122
10	1.1413	1.0000	-0.1413	1.0000	-0.1413

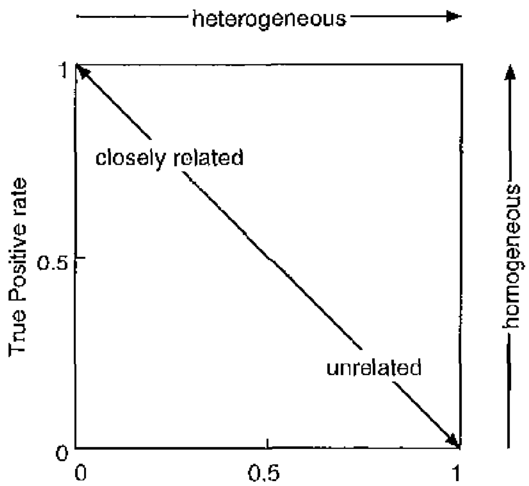

Figure 13. The relationship between the query and the derived rules on the ROC graph.

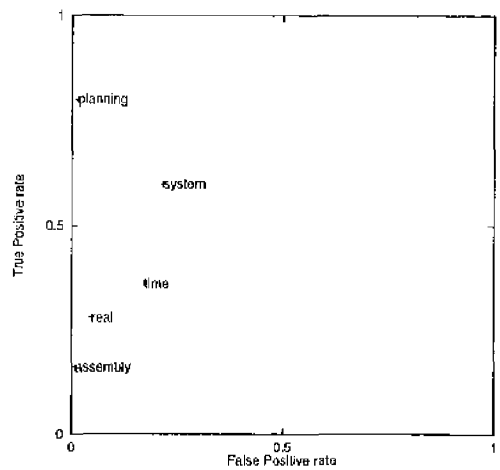

Figure 14. ROC map of associative keywords derived from a category 8 keyword "replanning".

mining association algorithm. The association algorithm uses confidence as the threshold and confidence is equivalent to TP in the ROC analysis, and confidence of a derived keyword is lower $Minconf$, it will be cut off and won't appear in the ROC graph. That is, the rule derived by the association algorithm appears to the FP direction on the ROC map, but the ROC map show another rule of the FP direction. And the FP direction shows heterogeneousness of keywords. For instance, in Figure 14 the keyword number of derived from "replanning" is few and it is easy to understand the meaning of the ROC map that "planning" is closely associative "replanning", and "system" has so higher TP but lower FP than the other keywords that "system" has many relationships to the other keywords and is less meaningful. In Figure 15 the keyword number of derived from "performance" is many and it is difficult to find out keywords on the ROC map. However by enlarging a part of the ROC map it is able to find out keywords as shown in Figure 16 , so "design" and "time" have the same TP but "design" has less relationship to the other keyword, hence closer. Therefore it is required for our navigation system to provide Graphical User Interface capable of enlargement and reduction, and capable of several visualization techniques such as clustering algorithm[4].

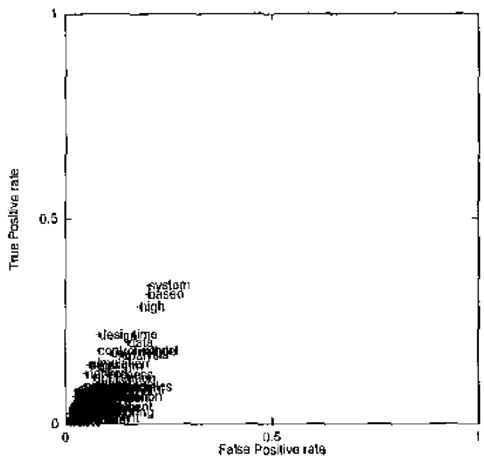

Figure 15. ROC map of associative keywords derived from a category 2 keyword "performance".

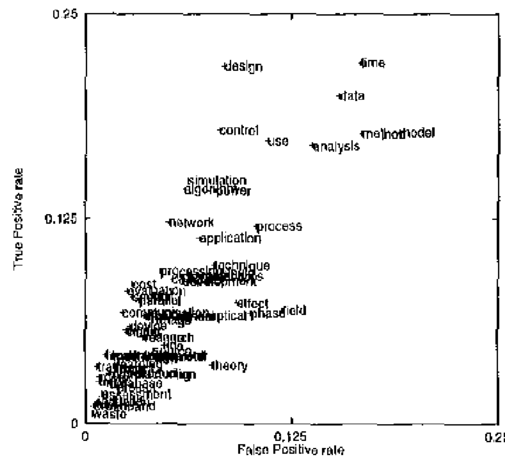

Figure 16. ROC map enlarged by the area of $0 \leq FP \leq 0.25$ and $0 \leq TP \leq 0.25$ in Figure 13.

6. Conclusion

The technique of data mining, such as association algorithm, becomes to be used widely in various fields. However it is too hard to specify the threshold for deriving rules in order to derive meaningful rules.

In this paper, we tried to evaluate the performance of rules derived by one of mining technologies based on the characters of ROC graphs considering the ROC convex hull method, and proposed the strategy to determine the threshold values in the basic algorithms for extended association rules. Moreover, we proposed the visualization method by the ROC map and shaw how to find out the relationship between a query and the derived keywords.

In the future, to derive more optimal and meaningful rules, we have to much more extend basic and ROC algorithms, which are proposed in this paper.

Acknowledgment

A part of this work is supported by the grant of Scientific Research (10780259, 08244103) from the Ministry of Education, Science, Sports and Culture of Japan. We are grateful to Nissho Iwai Infocom Co., Ltd. for the source programs of the full text search system "OpenText".

References

[1] C. Barber, D. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hull. Technical Report GCG53, University of Minnesota, 1993.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.

[3] J. Han, S. Nishio, H. Kawano, and W. Wei. Generalization-based data mining in object-oriented databases using an object cube model. *Data and Knowledge Engineering*, (25):55–97, 1998.

[4] M. Kawahara, H. Kawano, and T. Hasegawa. Evaluation of index based clustering in spatial database. *Proceedings of Advanced Database Symposium '96*, pages 79–86, 1996.

[5] M. Kawahara, H. Kawano, and T. Hasegawa. Data mining technologies for bibliographic navigation system. *Transactions of the IPSJ*, 39(4):878–887, 1998.

[6] M. Kawahara, H. Kawano, and T. Hasegawa. Implementation of bibliographic navigation system with text data mining. *Systems Science*, 3(24):101–118, 1998.

[7] H. Kawano. Mondou: Web search engine with textual data mining. *Proc. of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 402–405, 1997.

[8] H. Kawano, M. Kawahara, and H. T. The structure of text mining navigator for bibliographic search. *Symposium on Infology (In Japanese)*, pages 121–128, 1998.

[9] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of 3rd Int'l Conference on Knowledge Discovery and Data Mining (KDD-97)*, pages 43–48, 1997.

[10] R.Feldman. Practical text mining. In *Second Symposium on Principles of Data Mining and Knwoledge Discovery (PKDD-97)*, Nantes, France, 1998.

[11] R. Srikant and R. Agrawal. Mining generalized association rules. *Proceedings of the 21st VLDB*, pages 407–419, 1995.