データベースシステム 113－38
（1997. 7. 15）

# データマイニングのキーワード検索に対する応用

陳 漢雄　劉 野† 大保 信夫‡
つくば国際大学産業情報学科 †筑波大学工学研究科 ‡筑波大学電子情報工学系

概要

情報検索システムにおいて、文献のキーワード集合はその作者の研究内容と興味を表す。ユーザはこのキーワード空間について十分な知識を持てないため、最初から適当なキーワード問合せを与えることが困難である。本研究では関連ルール(Association Rule)の手法を用い、キーワード間に存在する関係を発見してユーザの問合せ修正を支援することを提案する。その実現として、対話的な支援システムを構築した。さらに、サポートと信頼性の高いものをルールとして採用する従来の方法では十分な絞り込みが困難であるなどの問題を解決する。一方、生成されるルールの数を押え、さらにルールの構造を明確に表現する目的で、Stem Ruleの概念を導入する。

# Information Retrieval by Keywords: A Data Mining Approach

Hanxiong Chen, Ye Liu† and Nobuo Ohbo‡
Department of Industrial Information, Tsukuba International University
† Doctoral Program in Engineering, University of Tsukuba
† Institute of Electronics and Information Science, University of Tsukuba
E-MAIL: {chx, liuye, ohbo}@dblab.is.tsukuba.ac.jp

概要

Data Mining is mainly used for discovering association rule in large databases, but few concentrate on screening the retrieval result. In this paper we present a query support method for document retrieval system by mining Association Rule between keywords from large document databases. A model is proposed to normalize the structure of mined Association Rules. We introduce a concept called "stem rule" from which all other association rules can be delivered. By this, the size of the rule base is considerably reduced. We build an interactive interface to aid user to refine their query. Empirical results confirm the screening effectiveness of our system.

## 1 はじめに

情報検索システムにおいて、ユーザが最初から適当な問合せを与えることは困難である。ユーザがデータの分布に対する完全な知識を持たないため、検索結果のサイズが大きすぎたり、小さすぎたりすることが発生する。ユーザは問合せの修正を繰り返し、ある程度のサイズの結果をチェックして、興味のある結果を出力させる。ユーザは自分の持っている知識を用いて問合せの修正を行うが、データベースの中には、ユーザの知らない知識が多く存在する。この部分の関係を利用しないと、ユーザの問合せは不完全なものとなることが多い。特に、文献検索システムの場合には文献のキーワード集合はその作者の研究内容と興味を表すため、このキーワードの空間についてユーザは十分な知識を持つことは困難であろう。

本研究では関連ルール (AssociationRule) の手法を用いて、キーワード間に存在する関係を発見し、これにより、ユーザの問合せの作成を支援することを試みる。関連ルールは項目 (Item) 間のサポートと信頼性により生成された関係である。文献検索の場合、ユーザの問合せになるキーワード集合に対して関連ルールを適用する。ユーザはこの関連ルールを参考にして、自分の問合せを修正する。

関連ルールを用いる通常のシステム [Agra93, Fayy96, Han95, Sava95, Srik96] では、サポートと信頼性の高いものがルールとして採用されてきた。しかし、文献検索においては、このような関連ルールを単純に適用すると次の各問題が生じる。

1. サポートの高いキーワードに基づく検索結果は出力サイズが大きくなり過ぎる傾向がある。
2. 信頼性の高いキーワードを問合せに追加しても出力の変化は少ない。
3. 最小サポートの制限で、頻度の低いキーワードは関連ルールとして認められない結果、問い合わせの十分な絞り込みが困難である。

本研究ではこれにかわって、極端にサポートが小さい場合を除き最大サポートと最大信頼性を用いて、ユーザにとって興味のある文献を検索できるように支援する。一方、最大サポートと最大信頼性を用いると膨大な数のルールが生成されるという問題に対処し、さらにルールの構造を明確に表現する目的で、Stem Rule 導入による手法を提案する。

次の章から、関連研究をサーベーし、我々の提案におけるルールの生成方法と Stem Rule の構造、有効性の実験、問合せインターフェースについて述べる。

### 1.1 関連研究

情報検索においては、問合せ拡張 (Query Expansion) [Xu96]、フィードバック (Relevance Feedback) [Alla95, Buck95, Chen94] などの問合せ支援方法がある。本節ではこれらの方法と我々の手法との関連及び相違点を明確にする。

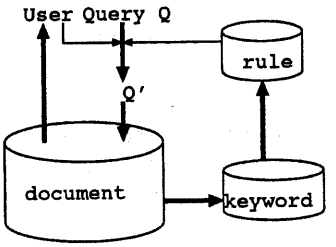
**フィードバック**は検索結果により、ユーザの問合せを修正する方法である。フィードバックはまず問合せと文献間の類似性を計算し、関連の文献を抽出する。次に抽出された関連文献のキーワードを利用し、問合せを修正する [Salt90]。

**問合せ拡張**は Association Hypothesis に基づいて、問合せを拡張する。即ち、類似性の高いキーワードを問合せに追加する。類似性は類似係数式により計算される [Peat91]。

この計算方法では、あるキーワードに対して、頻度の高いキーワードは類似性が高くなる。しかし、頻度の高いキーワードは文献の区別特性がよくないという欠点がある。これに対して、低い頻度のキーワードを用いて文献をクラスター化する方法もある。

以上の方法と我々の方法の共通点はキーワードの共起関係を利用して、関連のキーワード関係を発見するということである。我々の方法の特徴はキーワード間の関係を構造化し、問合せのパスが表示できる点にある。対話的な支援方式を用いて、ユーザとシステムの知識を統合する問合せが生成できる。

## 2 アプローチの概要




前にも述べたように、ユーザが文献データベースに対する検索を行なうとき、文献のキーワードに関する正確且つ詳細な知識を持つとは限らない。このため、問い合わせの初期段階では、漠然としたキーワード（集合）$Q$ を与えることが十分想定される。$Q$ 中のキーワードを全て含む文献をデータベースから検索することを仮定すると、膨大な文献数がユーザに返されてしまう。

この問題を解決するため、我々は関連ルールを利用するアプローチを提案する。このアプローチでは図で示したように、まず文献データベースからキーワードを抽出し、その中から絞り込み効果のあるキーワー

ドを関連ルールにより見つける。関連ルールはルールベースに格納し管理される。ユーザの問い合わせ $Q$ に対して、対話的にルールを適用し、検索結果を絞り込むようなキーワードを加える。この過程を繰り返し、最後にユーザの目的にあう問い合わせ $Q'$ を生成する。

## 3 Stem Rule によるアプローチ

### 3.1 用語と記号の定義

これからの議論に使われる記号を定義する。

定義 (操作 $\rho$) 対象文献の集合とキーワードの集合をそれぞれ $\mathcal{D}$ と $\mathcal{K}$ とする。文献 $d \in \mathcal{D}$ のキーワードを求める操作 $\rho : \mathcal{D} \to 2^{\mathcal{K}}$ を次のように定義する。

$\rho(d) = \{k | k \in \mathcal{K}$ かつ $k$ が $d$ のキーワードである $\}$

また、便宜上 $D \subset \mathcal{D}$ に対して、$\bigcup_{d \in D} \rho(d)$ のかわりに $\rho(D)$ と書く。

定義 (問い合わせ) $\mathcal{D}$ と $\mathcal{K}$ は同上とする。キーワードから $\mathcal{D}$ の文献を検索する問い合わせ $\sigma : 2^{\mathcal{K}} \to 2^{\mathcal{D}}$ は次のように定義される。

$Q \subset \mathcal{K}$ に対して

$$\sigma(Q) = \{\, d \mid d \in \mathcal{D}, \rho(d) \supseteq Q\}$$

即ち、$\sigma(Q)$ は $\mathcal{D}$ から $Q$ 中のキーワードを全て含むドキュメントを求める。

我々は関連ルールの適用について研究を進めているが、他のルールを扱わないので、今後は単に「ルール」と呼ぶ。ルール $r$ は適当な条件を満たさなければならないが、定義を述べるまでは次の形をいう。$r : X \Rightarrow Y,\ X, Y \subseteq \mathcal{K}$

$K \in \mathcal{K}$ の頻度を $Pr(K) \triangleq |\sigma(K)|/|\mathcal{D}|$ としたとき、ルール $r : X \Rightarrow Y$ のサポート (surport) と信頼性 (confidence) は次の式で計算される。

$$Spt(X \Rightarrow Y) = Pr(X \cup Y)$$
$$Cnf(X \Rightarrow Y) = Pr(X|Y) = Pr(X \cup Y)/Pr(X)$$

### 3.2 問題の形式化

従来のルールに関する関連研究では最小サポートと最小信頼性を用いてルールの数を押えている。しかし、前に挙げた問題点でも指摘したように、同じ方法を絞り込みを目的とする文献検索に適用すると、大量のルールを生成しかねないため、実用的ではない。そこで、本節ではまずルールの構造について形式化を試みる。

文献検索において、ルールは基本的にキーワード間の関係である。任意の $d \in \mathcal{D}$ を検索結果に含む問い合わせは $\rho(d)$ の子集合（つまり、$2^{\rho(d)}$ の要素）であるという点に着目すると、$DAG\ G = (N, E, \varphi)$ が我々の議論をカバーするベースになる。

ただし、

$$N = \bigcup_{d \in D} 2^{\rho(d)}$$

$$E \subseteq N \times N$$

$e =< n_1, n_2 > \in E \iff n_1 \subset n_2$

$\varphi$ は $E$ からルールへのマッピングであり、

$\varphi(< n_1, n_2 >)$ は $n_1 \Rightarrow (n_2 - n_1)$ というルールを返す。

$G$ に基づいて、関連ルールに対するサポート $Spt$ と信頼性 $Cnf$ を定義できるが、$e =< n_1, n_2 > \in E$ が決まれば対応するルールは一意に決まるので、$e$ に対して定義しても混乱は生じない。

簡単のため $\#n = |\sigma(n)|$ と書くと、明らかに

$$Spt(< n_2, n_1 >) = \frac{\#n_2}{|D|},\ Cnf(< n_2, n_1 >) = \frac{\#n_2}{\#n_1}$$

### 3.3 Stem Rule とその生成

従来のアルゴリズムを用いると、基本的に $O(2^{\mathcal{K}})$ 以上の計算量が必要で、理論的に同オーダののルールが生成可能である。*ad hoc* な閾値を使ってルール数を幾らか減らすことができるが、解析的な説明は全く不可能である。ルールの構造やルール間の関係も解明されない。これに対して、ルールの構造に基づいて導出関係を用いれば、すべてのルールを生成する代わりに「基本」となるルールのみを生成し、管理すれば良い。これが Stem Rule の基本的な考え方である。Stem Rule を導入するには、次のような基本条件を用いる。

$\theta_{s_l} < Spt(X \Rightarrow Y) < \theta_{s_u}$

$Cnf(X \Rightarrow Y) < \theta_c$

また、あるルール $r_1$ が基本条件を満たすとき、ルール $r_1$ も基本条件を満たすとき、 ルール $r_1$ はルール $r_1$ から導出可能であるという。

定義により次の各性質はほとんど自明である。

1. $\forall K \subset \mathcal{K}, 0 \leq |\sigma(K)| \leq |\mathcal{D}|,\ 0 \leq Pr(K) \leq 1$
2. $K \subset K' \subseteq \mathcal{K}$ ならば、$Pr(K) \geq Pr(K'), \sigma(K') \subseteq \sigma(K)$
3. $X \subseteq X', Y \subseteq Y', Spt(X \Rightarrow Y) < \theta_s$ ならば、$Spt(X \Rightarrow Y') < \theta_s$、$Spt(X' \Rightarrow Y) < \theta_s$、$Spt(X' \Rightarrow Y') < \theta_s$。
4. $Cnf(X \Rightarrow Y) < \theta_c$ ならば、$Cnf(X \Rightarrow Y') < \theta_c$。
5. $(X \Rightarrow Y)$ が基本条件を満たすとき、$(X \Rightarrow Y')$ も基本条件を満たす。

6. $(X \Rightarrow Y)$ が基本条件を満たすとき、$(X - \Delta X \Rightarrow Y \cup \Delta X)$ も基本条件を満たす。ただし、$\Delta X \subseteq \mathcal{K}$

定義 (Stem Rule) 基本条件と次の条件を満たすとき、ルール $X \Rightarrow Y$ を Stem Rule という。

$\exists d \in \mathcal{D},\ X \cup Y \subset \rho(d)$

次のアルゴリズムは文献集合 $\mathcal{D}$ から Stem Rule を生成する。従来のアルゴリズム ([Agra93]) のように、1-ItemSets, 2-ItemSets, ..., などを生成することから始めると、$n$-Itemset の生成には $O({}_nC_{|\mathcal{K}|})$ のキーワードの組合せをチェックしなければならない。一方、次のアルゴリズムは $d \in \mathcal{D}$ を対象に計算するので、組合せのチェックは一文献の平均キーワード集合に対してのみ行なう。

アルゴリズム（Stem Rule 生成）
FOREACH $d \in \mathcal{D}$
BEGIN
  $\rho(d)$ を求める
  FORALL $X \subset \rho(d)$
    IF $\theta_{s_l} < Spt(X) < \theta_{s_u}$ THEN
      $X \Rightarrow \phi$ を Cand に追加
END
WHILE Cand $\neq \phi$ DO
BEGIN
  Cand からルール $X \Rightarrow Y$ を取り除く
  FORALL $k \in X$
  BEGIN
   IF $(Cnf(X - \{k\} \Rightarrow Y \cup \{k\}) < \theta_c$
     且つ $X - \{k\} \Rightarrow Y \cup \{k\}$ が Stem Rule
     より導出できない) THEN
   $X - \{k\} \Rightarrow Y \cup \{k\}$ を Stem Rule に追加.
   OTHERWISE
   $X - \{k\} \Rightarrow Y \cup \{k\}$ を Cand に追加.
  END
  Cand から $X - \{k\} \Rightarrow Y \cup \{k\}$ より
  導出できるルールを削除。
END

### 3.4 ルールの適用

探索スペースは次の $DAG$ である:
$S = (N', E', \varphi')$
$S$ は $G$ の部分グラフであり、次の条件を満足する。
$N' \supseteq N \ \ (n \in N' \models \#n \leq \theta_{s_u})$
$E' \supseteq E \ \ (e = (n_1, n_2) \in E' \models \frac{\#n_2}{\#n_1} \leq \theta_c)$

ユーザの問い合わせ $Q$ に対して、ルールを適用しキーワードを加え、目的の問い合わせ $Q'$ にたどり着く考え方は $DAG\ S$ 中の任意のパス $L = (n_0 = Q, n_1, n_2, ...n_m = Q')$ の探索により実現される。対応するエッジのリストを $(e_1, e_2, ...e_m)$, $(e_i =< n_{i-1}, n_i >)$ とすると、

$$cnf(e_1) \times cnf(e_2) \times ...cnf(e_m)$$

$$= \frac{\#n_1}{\#n_0} \times \frac{\#n_2}{\#n_1} \times ... \frac{\#n_m}{\#n_{m-1}} = \frac{\#n_m}{\#n_0}$$

$$= Cnf(< n_m, n_0 >)$$

$cnf(e_i) < \theta_c$ から明らかに $cnf(< n_m, n_0) < \theta_c$ が成り立ち、つまり、次の結論が自明になる。

Property 1. $S$ に $L = (n_0 = Q, n_1, n_2, ...n_m = Q')$ が存在するための必要条件は $< n_0, n_m > \in N'$.

Property 2. $S$ に $L = (n_0 = Q, n_1, n_2, ...n_m = Q')$ が存在するための十分条件は

$Cnf(< n_{i-1}, n_i >) \leq \sqrt[n]{\frac{\#n_m}{\#n_1}}$.

アルゴリズム（ルール適用）
入力：ユーザ問い合わせ $Q = Q_0$
1. FORALL Stem Rule $X \Rightarrow Y$
  IF $X \supset Q_i$ THEN
  $Q_{i+1} = X \cup Y$

2. ユーザに対して、$Q_{i+1}$ を示す。

3. ユーザからストップの指令がなければ 1 にもどる

## 4 実験

### 4.1 実験用データ

我々は電気工学分野の 4 万件の文献 ($|\mathcal{D}| = 40k$) を対象として実験を行なった。この 4 万件の文献に含まれるキーワード数は 16717 個ある ($|\mathcal{K}| = \rho(\mathcal{D}) = 16717$)。まず最初に文献 $d(\in \mathcal{D})$ のキーワード数 ($|\rho(d)|$) の分布を統計し、結果を表 1 に示す。この表から、ほとんどの文献のキーワード数は 15 以内であることがわかった。キーワード ($k \in \mathcal{K}$) のサポート ($spt(k)$) の分布は表 2 に示す。サポート (Spt) の単位は文献の数である。三分の二強のキーワードのサポートは 10 以下 (0.025%) である。このようなキーワードを使うと、結果が分散し過ぎて、ユーザが候補の選択に困るのでサポートは 10 以上に限定する ($\theta_{s_l} = 0.025\%$)。

このデータに対して、ルールの生成と管理及び問合せへの支援などの実験を行なう。

| $\|\rho(d)\|$ | 1-5 | 6-10 | 11-15 | 16-20 | 21-25 | 26-30 | 31-33 | Total |
|---|---|---|---|---|---|---|---|---|
| $\|D\|$ | 1935 | 25902 | 11029 | 1029 | 91 | 13 | 1 | 40000 |

表 1. 一文献のキーワード数の度数分布

| Spt | 1-10 | 11-100 | 101-200 | 201-300 | 301-500 | 501-1000 | 1001-2100 | Total |
|---|---|---|---|---|---|---|---|---|
| $\|K\|$ | 11119 | 4772 | 524 | 156 | 94 | 47 | 5 | 16717 |

表 2. キーワードのサポートについての度数分布

| $n$ Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\|n-$Itemsets$\|$ | 5886 | 16499 | 6389 | 2155 | 1244 | 886 | 490 | 177 | 36 | 3 | 33765 |

表 3. $Spt \geq \theta_{s_l}$(10 文献) の Itemsets のサイズについての統計

| $\|Y\|$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| All Rule | 51548 | 42029 | 37717 | 34207 | 22512 | 9828 | 2721 | 433 | 30 | 201025 |
| Stem Rule | 51548 | 4957 | 18008 | 2092 | 5 | 0 | 0 | 0 | 0 | 76610 |

表 4. 生成された全てのルールと Stem Rule 数の比較

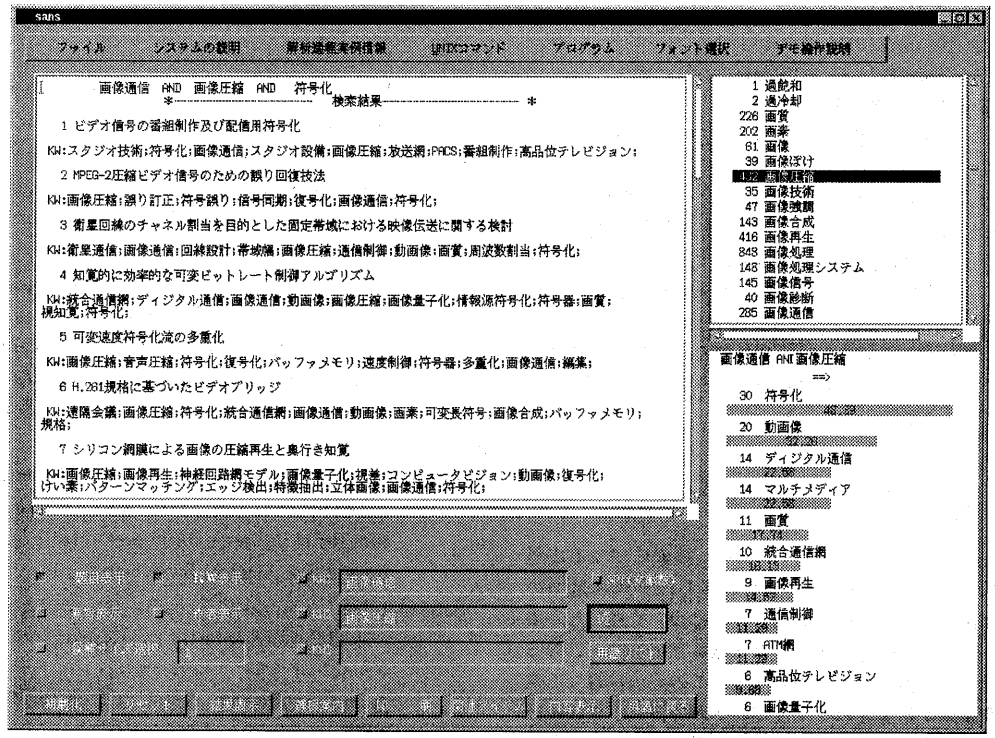



図. インタフェース

## 4.2 ルールの構造

$Spt(K) \geq \theta_{s_l}$ $(K \subset \mathcal{K})$ の条件での ItemSets サイズの統計結果は表 3 に示す。表 4 に $\theta_{s_l} = 10$、$\theta_c = 60\%$ の場合の生成ルール数を示す。ここで $|Y|$ はルールの右辺のキーワード数である。この表で分かるように、最初のスキャン $(Y = 1)$ は計算が非常に簡単で、両方とも 5 万あまりのルールを生成した。しかし、従来の方法はさらに複雑な組合せチェックを行なって 15 万前後ものルールを生成したのに対して、我々が提案した Stem Rule を用いたとき 2 万あまりのルールしか生成しない。

### 4.3 問合せへの支援

図は問合せへの支援の例を示す。ユーザが $Q_0$={画像通信 (spt=285), 画像圧縮 (spt=432)} を与えると、システムは次のルールから図の参考リストを生成する。

{ 画像通信, 画像圧縮 }
⇒ { 符号化 } (spt=30,cnf=48.4%)
{ 画像通信, 画像圧縮 }
⇒ { 動画像 } (spt=20,cnf=32.3%)
{ 画像通信, 画像圧縮 }
⇒ { マルチメディア } (spt=14,cnf=22.6%)
{ 画像通信, 画像圧縮 }
⇒ { ディジタル通信 } (spt=14,cnf=22.6%)
{ 画像通信, 画像圧縮 }
⇒ { 統合通信網 } (spt=10,cnf=16.1%)
{ 画像通信, 画像圧縮, 動画像 }
⇒ { 符号化 } (spt=12,cnf=60%)
{ 画像通信, 画像圧縮, 符号化 }
⇒ { 動画像 } (spt=12,cnf=40%)
*{ 画像通信, 画像圧縮 }
⇒ { 符号化, 動画像 } (spt=12,cnf=19.4%)

(* は Stem Rule から導出されたルールである。)
ユーザは関連のキーワード候補リストを参考して、例えば「符号化」を入力あるいはクリックによって選び、$Q$ を補完し $Q_1$={画像通信 (spt=285), 画像圧縮 (spt=432), 符号化 (spt=692)} を作成する。これで満足すればこの $Q_1$ を問い合わせとして文献を検索すればいいが、さらに絞り込みを行ないたい時は同じようにルールを適用すれば良い。

## 5 まとめ

データマイニングのキーワード検索に対する応用として、stem rule の概念を導入し、それに基づいて関連ルールの生成、管理と適用について述べた。また、電気工学分野の4万件の文献を利用して、この方法の効果を実験により検証した。この方法に対して次の問題点が指摘される。サポートが非常に小さなルールを適用することは、ルール適用アルゴリズムの2においてユーザに表示する $Q_i$ を大量に発生させる。

今後の検討課題として、シソーラスを参考に、「総称的な」概念の導入が考えられる。ここで、$X, Y \in \mathcal{K}$ に対して、$\sigma(X) \supset \sigma(Y)$ のとき X が Y より総称的であるという。$Y_1, Y_2$ にたいし、総称的な $Y$ が存在するとき、ルール $X \Rightarrow Y_1$, $X \Rightarrow Y_2$ を $X \Rightarrow Y$ に減らすことができる。

## 参考文献


[Agra93] R.Agarawal, T.Imielinski and A.Swami. Mining Association Rules between Sets of Items in Large Databases. *ACM SIGMOD'93*, pp.207-216.

[Alla95] J.Allan. Relevance Feedback With Too Much Data. *ACM SIGIR'95*, pp.337-343.

[Buck95] C.Buckley and G.Salton. Optimization of Relevance Feedback Weights. *ACM SIGIR'95*, pp.351-357.

[Chen94] C.M.Chen and N.Roussopoulos. Adaptive Selectivity Estimation Using Query Feedback. *ACM SIGMOD'94*, pp.161-172.

[Fayy96] U.Fayyad, G.Piatestsky & P.Smyth. From Data Mining to Knowledge Discovery in Databases. *3st Knowledge Discovery and Data Mining*, pp.37-53, Califinia, USA, 1996.

[Han95] J.Han and Y.Fu. Discovery of Multiple-Level Association Rules from Large Databases *21st VLDB*, pp.420-431, Zurich, Swizerland, 1995.

[Peat91] H.J.Peat and P.Willett. The Limitations of Term Co-Occurrence Data for Data for Query Expansion in Document Retrival Systems. *Jounal of The American Society for Information Science*, vol.42(5), pp.378-383, 1991.

[Sava95] A.Savasere, E.Omiecinski and S.Navathe. An Efficient Algorithm for Mining Association Rules in Large Databases. *21st VLDB*, pp.432-444, Zurich, Swizerland, 1995.

[Salt90] G.Salton annd C.Buckley. Improving Retrieval Performance By Relevance Feedback. *Jounal of The American Society for Information Science*, vol.41(4), pp.288-297, 1990.

[Srik96] R.Srikant and R.Agrawal. Mining Quantitative Association Rules in Large Relational Tables. *ACM SIGMOD'96*, pp.1-12.

[Xu96] Jinxi Xu and W.Bruce Croft. Query Expansion Using Local and Global Document Analysis. *ACM SIGIR'96*, pp.4-11.