Association Rule の文献検索への応用

5K－2　出沢 信雄　脇山 賢一†　大保 信夫††

筑波大学 理工学研究科 † 筑波大学 工学研究科 †† 筑波大学 電子・情報工学系

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る問題が発生する。ユーザは問合せの修正を繰り返し、結果のサイズを利用して、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 (Association Rule) 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膨大になる傾向がある。
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ため、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、上で述べた3つの問題を考慮して、サポートと信頼性の値がある制限の中にあるような、より文献検索に適したルールである Stem Rule[1] の導入による手法を提案する。

Application of Association Rules to Document Retrieval
Nobuo Idezawa, Kenichi Wakiyama†, Nobuo Ohbo††
Master's Program in Sci. and Eng., Univ. of Tsukuba
†Doctoral Program in Eng., Univ. of Tsukuba
††Institute of Info. Sci. and Elec., Univ. of Tsukuba

2 アプローチの概要

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か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文献検索システムを提案する。このシステムでは、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。しかし、見つかった関連ルールが非常に大きかった場合も十分に考えられる。そこで以下の Stem Rule の概念を導入し、提示するキーワードの数をユーザの負担に耐える程度に抑える。この過程を繰り返し、最後にユーザの目的にそう問い合わせ Q' を生成する。

3 Stem Rule

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 ρ) 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。

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、我々はルールの構造に着目して考慮すべきルールの数を抑制する方法を持用した。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、$DAG\ G = (N, E, \varphi)$ が我々の議論のベースになる。

ただし、

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

$$E \subseteq N \times N$$

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

φ 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。

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 $O(2^{\mathcal{K}})$ の計算量が必要で、理論的に同オーダのルールが生成されてしまう。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かわりに、前で述べた従来の関連ルールを単純に適用した時の3つの問題を考慮して「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 が基本条件を満たし、ルール r_2 も基本条件を満たすとき、ルール r_1 はルール r_2 から導出可能であるという。

定義 (Stem Rule) 基本条件を満たすとき、ルール $\varphi(< n_1, n_2 >)$ をStem Ruleという。
ただし、$|n_2 - n_1| = 1$

[1]では、電気工学分野の4万件の文献 ($|\mathcal{D}|$=40000) を対象に実験を行なっている。なお、これらの文献には16717個のキーワード ($|\mathcal{K}|$=16717) が含まれている。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、次に、キーワード ($k \in \mathcal{K}$) のサポート (spt(k)) の分布を求め、この結果を元に θ_{s_l} を決めた。これに、θ_{s_l} の条件を加えルールを生成したところ約20万のルールを生成した。しかし、Stem Ruleを用いたときは約8万のルールしか生成していない。

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。
$N' = \{n | n \in N \wedge \frac{|\sigma(n)|}{|\mathcal{D}|} \leq \theta_{s_u}\}$
$E' = \{(n_1, n_2) | (n_1, n_2) \in E \wedge \frac{|\sigma(n_2)|}{|\sigma(n_1)|} \leq \theta_c)\}$

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、

$$cnf(e_1) \times cnf(e_2) \times ...cnf(e_m)$$
$$= \frac{|\sigma(n_1)|}{|\sigma(n_0)|} \times \frac{|\sigma(n_2)|}{|\sigma(n_1)|} \times ... \frac{|\sigma(n_m)|}{|\sigma(n_{m-1})|} = \frac{|\sigma(n_m)|}{|\sigma(n_0)|}$$
$$= cnf(< n_m, n_0 >)$$

$cnf(e_i) < \theta_c$ 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 $< n_0, n_m > \in E'$

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

$$cnf(< n_{i-1}, n_i >) \leq \sqrt[m]{\frac{|\sigma(n_m)|}{|\sigma(n_1)|}}$$

4 おわりに

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。

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 に対し、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

[1] Ye Liu, Hanxiong Chen, Jeffrey Yu and Nobuo Ohbo. Using Stem Rules to Refine Document Retrieval Queries. *Inte'l Conf. on Flexible Query Answering System (FQAS'98)*, Roskilde, Denmark. also in *LNAI No.1495, pp.249-260. Springer-Verlag.* 1998