Association Rules を利用した文献検索システムの研究

5K－1

脇山 賢一　出沢 信雄†　大保 信夫‡

筑波大学工学研究科 †筑波大学理工学研究科 ‡筑波大学電子・情報工学系

1　はじめに

情報検索において、ユーザが最初から適当な問い合わせを与えることは困難である。そのため、ユーザからの問い合わせを詳細化していくシステムに関する研究が数多く行なわれている[1]。本稿では、文献データベースから抽出される知識をもとに、ユーザが対話的に問い合わせ修正を繰り返すことで、ユーザの意図する問い合わせを生成し、最適な文献を得ることを目的としたシステムを提案する。文献データベースからの知識抽出法としては、文献とそれに含まれるキーワードの関係に着目する。このキーワードの集合の中からトランザクションデータベースにおける関係ルールと類似の関連ルール（Association Rules）を求める。ここでの関連ルールとは、キーワードの共出現性に基づくもので、従来から二つのキーワードの共出現を利用したシステムは多く研究されているが、関連ルールにおける $X \Rightarrow Y$ と $Y \Rightarrow X$ の非対称性が問い合わせ修正においては極めて有効である。しかし、単純な関係ルールの適用では考慮すべきキーワード集合が極めて大きなものになるため、それを実用的レベルまで削減する目的でカバーの概念を導入する。

2　文献検索におけるシステム

本研究における文献検索システムとは、キーワード集合が与えられた時、それらのキーワードをすべて含む文献集合を検索する問題のことである。ユーザにより与えられたキーワード集合に何らかのキーワードを加えることにより出力文献集合は縮小する。この時、加えるキーワードを適切に選択することによりユーザの意図に則した文献結果集合を得ることができる。また、加えられるキーワードはユーザの検索意図をよく表現し、かつ検索空間である文献データベースの性質を反映したものであるべきである。

このことを実現するために、文献データベースの性質を反映させる手段として関連ルール（Association Rules）の手法を採用する。ここでの関連ルールとは、キーワード集合 $\mathcal{K}$ の部分集合 $X \subseteq \mathcal{K}, Y \subseteq \mathcal{K}$ に対し

$$X \Rightarrow Y, C$$

で表される。これは、キーワード集合 X をすべて含む文献集合は、C% の割合でキーワード Y をすべて含むことを意味している。また、キーワード集合 X と Y をすべて含む文献数の全文献数に対する割合をサポート（Spt）と呼び、C の値を信頼性（Cnf）と呼ぶ。また、従来の最小サポートと最小信頼性に基づく関連ルールとは異なり、以下のような条件でルール生成の数を抑えている。

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

ユーザの意図は、関連ルールにより提示されたキーワード集合の中から選択することで示される。

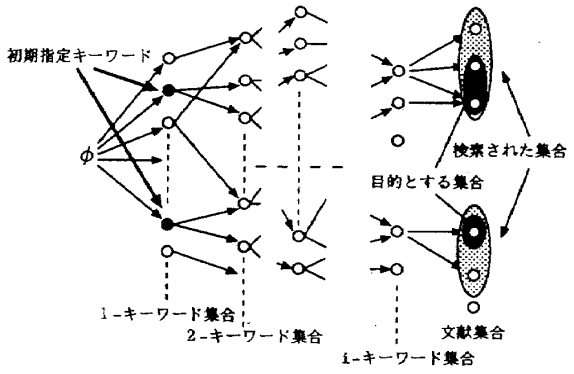

図 1: キーワード空間

いま図1で示されたグラフの各ノードを i-キーワード集合とし i 個のキーワードから成る集合を表すとする。またこれに、そのキーワード集合を含む文献集合を対応させる。ノードとノードの間の矢印は直接的な関連ルール（Stem Rule）を示している。この時、ユーザの求めたい文献集合を斜線部分で示された部分とし、黒点をユーザの初期指定のキーワード集合とする。

この時、我々の目的とするシステムは、この黒点から文献集合へ効率良くユーザを導くものである。ユーザによってキーワード集合 X が指定された時、上記グラフ中のアーク $X \Rightarrow Y$ に対応して、候補となるキーワード集合 $Y - X$ と予想される出力文献の数をユーザに提示する。予想される出力文献の数は、キーワード集合 X を含んでいる文献数を N とすると、関連ルール $\{X \Rightarrow Y, C\}$ の信頼性 C を用いることで

$$N \times \frac{C}{100}$$

Document Retrieval System with Association Rules
Kenichi WAKIYAMA, Nobuo IDEZAWA†, Nobuo OHBO‡
Doctoral Program in Eng., Univ. of Tsukuba
†Master's Program in Sci. and Eng., Univ. of Tsukuba
‡ Institute of Info. Sci. and Elec., Univ. of Tsukuba

で計算される。これにより、次に選択するキーワードによって得られる文献の数が分かる。このような提示を繰り返すことでユーザは、最終的に表示される文献数を確認する。

3 関連ルールの問題点

上のような文献検索システムのプロトタイプを実装し、電気工学分野の4万件の文献 ($|\mathcal{D}|$=40k) を対象として実験 [1] を行なった。この4万件の文献から抽出されたキーワード数は16717個である ($|\mathcal{K}|$=16717)。一つの文献あたりのキーワード数は、ほとんど15個以下であった。また、関連ルール生成の際の基本条件として $\theta_{s_l} = 10, \theta_c = 60$ を用いて、関連ルールを生成した結果、約20万個の関連ルールが生成された。
このように、関連ルールとして莫大な数のルールが生成される。その結果、ユーザの指定したキーワード集合 X に対し、候補として提示されるキーワード集合 $Y - X$ の数は極めて多いものとなる。

4 カバー

本研究では、ユーザよってキーワード集合 k が与えられた時、ユーザに提示する候補となるキーワードを選んで提示することを考える。

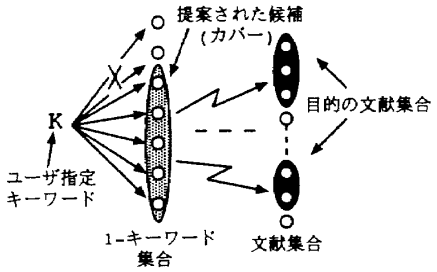

図 2: カバーの例

この時、提示される候補のキーワード集合の満たすべき制約条件としては、候補キーワードを含む文献集合全体が、ユーザの指定した問い合わせのキーワード集合 K を含む文献集合すべてを含んでいなければならないということである。もし、そうでなければキーワード集合 K が与えられた時、検索されるはずの文献の中で新しいキーワードを加えることで、検索できなくなってしまう文献が生じてしまうからである。

$\mathcal{D}$ と $\mathcal{K}$ を同上とし文献集合 D のうち、キーワード k を含む文献集合を d_k とする。その時、$D \subseteq \mathcal{D}, K \subseteq \mathcal{K}$ に対して

$$D \subseteq \bigcup_{k \in K} d_k$$

が成り立つ時、K が D をカバーする（或いは、K が D のカバーである）という。また、$K' \subset K$ なる D のカバー K' が存在しない時、K が D の極小カバーである。同じように、次の式が成り立つ時、K が $K'(\subseteq \mathcal{K})$ をカバーするという。

$$\bigcup_{k \in K'} d_k \subseteq \bigcup_{k \in K} d_k$$

なお、便宜上 $\{(K \cup K') | \ K \Rightarrow K' \in R\}$ が D をカバーする時、ルール集合 R が D をカバーするという。ユーザには、この極小カバーの一つを提示することで提示されるキーワードの数を削減する。

5 文献検索の概要

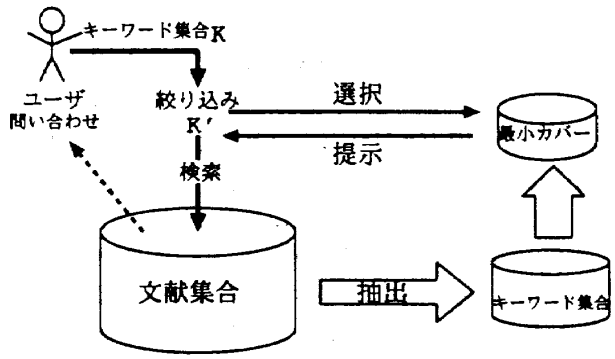

図 3: 検索の概要

図3は、ユーザからキーワード集合 K が与えられてから、検索されるまでの様子を表している。ユーザからキーワード集合 K が与えられると、まず、キーワード集合 K の極小カバーの一つをユーザに提示する。ユーザは予想文献の数をもとにキーワードを絞り込むことを繰り返し、文献検索を行なう。

6 まとめ

文献検索システムにおける関連ルールの適用を考え、その応用としてカバーの概念ををを導入した。その際、提示すべき極小カバーの選択方法などを考えることによってユーザにとって最適であるべき文献検索の問題を考えていく必要がある。

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, of Springer-Verlag* 1998.