異種データベースからのデータ発掘による検索支援

2AD－1

河野 浩之

京都大学大学院工学研究科

1. はじめに

既存の異種データベース研究の枠組みを越えた考察をwebシステムは必要としているのは，多数のユーザの要求が様々な角度から社会性を帯びて顕在化するシステムとなっているからである．つまり，複数データベースを通信ネットワークにより連携するシステム構成を考える上で，非常に重要な切口をwebは与えている．今後，効率的な情報共有を可能とするための情報システム基盤を整備する上で，webシステムに対するウェブ発掘(web mining)[4]研究は，異種データベースの一つのリファレンスモデルとなりうる重要な足掛かりである．

本稿では，異種データベース検索支援に対するデータマイニング技術を課題として取上げ，特に，異種データベース環境の典型例となるwebシステムに対するデータマイニング技術[5]を用いた実装までの過程を論じる．

2. 異種DBマイニングの役割

過去，異種データベース統合技術として議論されたシステム構成は，異種分散データベース管理システム(HDDBMS, heterogeneous distributed database management system)のような複数のレガシーデータベース(Legacy database)の相互運用を可能とする構成技術であった．

これを，webシステム上でのシステムの連携として考えると，WebSQL (http://www.cs.toronto.edu/~websql)研究[2]のように，各種リソースを関係モデルやオブジェクト指向モデルとして捉えた枠組みに相当する．

しかし，現在のデータベース統合技術における問題点は，データマイニング技術の応用例として頻繁にとりあげられるシステム構成の一例であるデータウェアハウス(data warehouse)において典型的に生じる，異質な複数のデータベースを協調させる枠組みである．特に，統合すべき異種データベースが同一組織内で構築されたものでない場合，情報システムの利用形態に応じて，数多くの多様な情報システムを緩やかに連携させることを想定する必要がある．

このような問題に対して，構造化されていない半構造データを含む異種情報システムに蓄えられたデータの統合を目指したプロジェクトとして，スタンフォード大学におけるTSIMMIS (The Stanford-IBM Manager of Multiple Information Sources) (http://www-db.stanford.edu/tsimmis/tsimmis.html)[6]などが重要である．

しかしながら，メディエータやラッパー等を用いた手法によって，独立して構築された異種情報システムを統合するには，複数データベースを相互に透過的に検索するような質の高い知識が必要であることから，柔らかな連携システムを構成するには困難な点が多い．

つまり，図1に示したような，質の異なる情報空間を効率良く検索するためには，単一のデータベース検索よりも優れた検索式記述が要求され，有効性の高い検索式を記述するために深い領域知識が必要とされるからである．

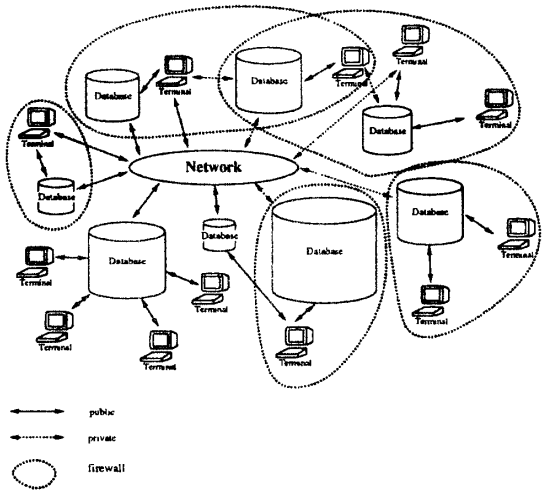

図1: 異種データベース情報空間

そこで，我々は，数多くのシステムと連携する知識記述の基礎として，データマイニングアルゴリズムの代表例であるApriori[1]に代表されるアルゴリズムを拡張し，検索エージェント間で導出された相関ルールを交換することを試みる．なお，各々の情報システムにエージェント(agent)層をもうけ，システム間でKQML (Knowledge Query and Manipulation Language, "http://www.cs.umbc.edu/kqml/")[8]を用いて異種性の吸収を試みる．

A Framework of Intelligent Search System Using Knowledge Discovery in Heterogeneous Databases
Hiroyuki Kawano
Kyoto University
Kyoto 606-01, Japan

3. 異種データベースとしてのWeb検索

web空間を構成する情報資源は，全く独立した組織によって，本質的な自由さをもって作成されており，キーワード空間やデータ相互のリンク構造などの制約を想定することは全く不可能である．

そこで，我々は，「問答」の枠組みを用いて，対話性の高い検索支援を与えるRCAAUシステム (http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/) による実験を試みている．特に，入力キーワードに関係するルールをマイニング処理によって求め，ユーザの検索要求の近傍のキーワード空間の構造を関連語の知識として与える検索支援が有効であることを，多数のユーザによる1,000,000件を越える検索要求を分析することで検証している．

さらに，異種データベース構成を考える上で，webのみではなく，電子ニュースシステムや電子メールなどの文書データに対する検索システムを構築し，それぞれのシステムから相関性などのルールを導くデータマイニング機能を実装した．そして，初期検索式に対してルール導出を行うだけではなく，KQMLにより複数システムを連携させることにより，異質な検索空間の特性を把握した検索式を生成し，図2に示したようなシステム構成により検索環境改善の可能性を検証している．

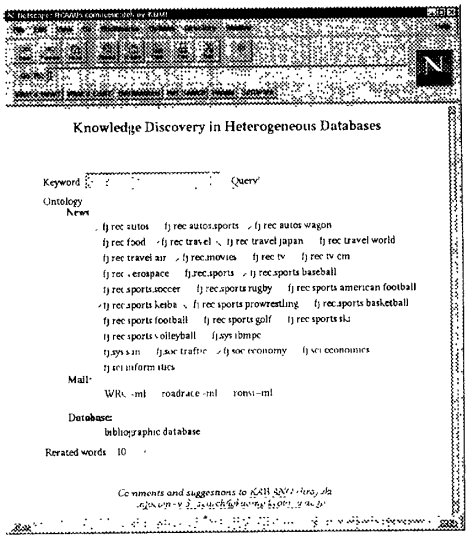

図2: 分散環境でのプロトタイプシステム

例えば，検索語として“学習”を与えた場合，webからは，「センター，基，生涯，特集，関，研究，基礎，学力，教育」が，国立国会図書館の雑誌記事索引データベースからは，「特集，問題，地球，開発，エネルギー，都市，企業，教育，技術」が関連語として提示される．このようなルールを導出することによって，異種データベースから導出された相関ルールによる検索支援システムの評価を行う[7]．

さらに，データベースに関するセキュリティ面[3]からも，単純な全文検索結果ではなく，ルール形式の情報交換は有効であると考えられる．ただし，たとえルール形式であっても，多数の問合わせにより導かれるルール集合に対するセキュリティなどに関する議論は残されている．

4. おわりに

複数の情報源を通信ネットワークによって効果的に統合し，より質の高い知識を得るためには，複数のデータベースを連携させて導出されるルールを如何にして統合するかという技術が必要となっている．既に，単純な構造をもつ「複数のWeb検索エンジン」による統合の実装を行う上で，異種データベース統合において解決すべき数多くの問題点が明らかになっており，将来の柔らかな協調型検索システム構築のために解決すべき課題は非常に多い．

今後も特定の観点からデータが整理されている種々の情報システムが構築されると思われるが，それらのデータを別の軸から再統合し整理し直す基礎技術，すなわちデータの組織化を支援するアルゴリズムが重要な役割を果たすと考えられる．

謝辞

本稿の一部は，文部省科学研究費重点領域研究(1)(08244103)のもとでの研究成果となっている．

参考文献

[1] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc. of the 20th International Conference on Very Large Data Bases, Santiago, Chile, pp.487–489, 1994.

[2] G. Arocena, A. Mendelzon, and G. Mihaila, “Applications of a Web Query language,” Proc. of the 6th International WWW Conference, Santa Clara, California, April 1997.

[3] M. S. Chen, J. Han and P. S. Yu, “Data Mining: An Overview from a Database Perspective,” IEEE Transactions on Knowledge and Data Engineering, vol.8, no.6, pp.866-883, 1996.

[4] O. Etzioni, “The World-Wide Web: Quagmire or Gold Mine?,” Communications of the ACM, Vol.39, No.11, pp. 65–68, Nov 1996.

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

[6] H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom, “Integrating and Accessing Heterogeneous Information Sources in TSIMMIS,” Proc. of the AAAI Symposium on Information Gathering, pp. 61-64, Stanford, California, March 1995.

[7] 川原稔，河野浩之，長谷川利治，“WWWデータ資源検索におけるデータマイニング手法，” 情処研報97-DBS-112, pp.33–40, 1997.

[8] J. Mayfield, Y. Labrou and T. Finin, “Evaluation of KQML as an Agent Communication Language,” Proceedings of the 1995 Workshop on Agent Theories, Architectures, and Languages, Springer-Verlag, 1996.