DEWS2004 I-5-01

異種のデータベース統合による複数情報源の統合的な閲覧環境の実現

佐藤　大介† 　河本　稼†† 　清木　康†††

† 慶應義塾大学総合政策学部　〒 255–8520 神奈川県藤沢市遠藤 5322
†† 慶應義塾大学政策・メディア研究科　〒 255–8520 神奈川県藤沢市遠藤 5322
††† 慶應義塾大学環境情報部　〒 255–8520 神奈川県藤沢市遠藤 5322
E-mail: †{s00450ds,k12u,kiyoki}@sfc.keio.ac.jp

あらまし　本稿では，報道記事や法令などの異種のデータベース群を対象とした統合的な参照および閲覧環境を実現し，データベース間に散在する関連情報を動的に連結・参照するマルチデータベース環境の実現方式を示す．異種データベース群をマルチデータベース環境に連結・統合を行うことにより，利用者は異種の情報源間に存在する関連情報を動的に参照し，統合的な閲覧を行うことが可能となる．本方式の特徴は，利用者が閲覧する記事などのドキュメントの特徴を解析し，統合の対象する専門データベースを動的に選択する点にある．具体的には，本方式の特徴は，記事などのドキュメント群を，その特徴によりジャンルごとにカテゴリに分ける．そして，記事に関連する企業情報や法令のデータベース，書籍のデータベースなど，記事に関連するデータベースを動的に選択および接続し，記事の関連情報としてそれらデータベースのコンテンツを選択し，記事に並置して提示することを実現する．
キーワード　情報統合，異種 DB，情報検索，Web とインターネット，テキスト DB

Implementation of Integrative Browsing Environment for Heterogeneous Information Resources

Daisuke SATO†, Minoru KAWAMOTO††, and Yasushi KIYOKI†††

† Faculty of Policy Management, Keio University　Endo 5322, Fujisawa, Kanagawa, 255–8520 Japan
†† Graduate School of Media and Governance, Keio University　Endo 5322, Fujisawa, Kanagawa, 255–8520 Japan
††† Faculty of Environmental Information, Keio University　Endo 5322, Fujisawa, Kanagawa, 255–8520 Japan
E-mail: †{s00450ds,k12u,kiyoki}@sfc.keio.ac.jp

Abstract　In this paper, we present a multidatabase environment for browsing heterogeneous document databases by computing inter-relationships among documents. By connecting and integrating heterogeneous databases in the multidatabase environment, users can view inter-relationships among heterogeneous information sources, and browse integrated documents. The feature of this method is to analyze features of documents, news articles for instance, and to dynamically select appropriate databases according to requests by using the results of the document analysis. This method provides a capability to categorize documents into genre according to the features of the documents, and to dynamically integrate heterogeneous document databases, such as commercial, law and publication databases, by computing the inter-relationships with others for browsing. By using this method, users can compute inter-relationships among heterogeneous document databases to dynamically select valid databases.

1. はじめに

ネットワークの広域化，メディアコンテンツの拡大は，利用者の知識獲得における利便性を大きく向上させている．データベースアクセスの増加は，利用者にとって新たな知識発見の機会提供に大きく寄与し，その重要性は今後ますます増大していく．特に，専門データベース群として，法令関係情報，書籍情報などのデータベース群は，専門家に限らず幅広いユーザが，具体的な対象を扱う学習や知識発見の場として重要なものであり，学際的な情報探索統合活動をより活性化させるものになる．

ユーザがそれらのデータベース群を利用する場合，日頃身近に触れている情報と専門的な情報源群には前提となる専門知識において大きな格差があり困難を伴う。対象として扱う情報の専門性が高い場合、対象となる専門分野の専門知識なしには獲

得した情報について理解することが難しく、効率が悪い。学問の専門性の高い場合に専門的な用語を用いて情報を獲得するためには，それぞれの専門データベースから関連する単語により検索しなければならない．この方法では，知識獲得が難しく，また効率も悪い．

本稿では，報道記事などの一般的なドキュメントデータベース群，および，法令，政治，経済等の専門的なドキュメントデータベース群を，マルチデータベース環境において統合し，また，ローカルデータベース群のメタレベル (抽象層，上位層) においてドキュメント解析の手法を適用することで，それら異種のドキュメントデータベース群を対象とした統合的な閲覧環境の実現方式を示す．異種データベースをマルチデータベース環境に連結・統合を行うことにより，利用者は異種の情報源間に存在する関連情報を動的に参照し，統合的な閲覧を行うことが可能となる．

本方式の特徴は，利用者が閲覧する記事などのドキュメントの特徴を解析し，統合の対象とするデータベースを選択する点にある．具体的には，本方式は，記事などのドキュメント群を，その特徴によりジャンルごとにカテゴリに分ける．記事に関連する企業情報や法令のデータベース，書籍のデータベースなど，記事に関連するデータベースを動的に選択および接続し，記事の関連情報としてそれらデータベースのコンテンツを選択し，記事に並置して提示することを実現する．

本方式により，記事などに含まれる平易な言葉から関連する複数の専門データベースにおける情報を獲得することが可能となり，高い専門性のあるデータベースを対象として，一般記事などの専門知識を必要とすることなく知識として獲得できる情報を用いて、専門性の高い情報を獲得する環境を実現する．本方式は、新たな知識発見の機会をより多く提供し，利用者にとって情報獲得活動をより活性化させるという視点を実現する方式として位置付けることができる．

異種の複数情報源を統合するためには，各専門データベースにおける情報群を特徴付ける必要がある．本方式では，異種の複数情報源の統合的な閲覧環境を実現するために，次の 3 機能を設定する．

(1)　各専門データベースにおけるデータ群を特徴付け特徴語を抽出するための機能

(2)　検索統合の際にどの専門データベースにアクセスするかを選別するために，ドキュメントデータベースと専門データベースの関連性を計量する機能

(3)　学問の専門性に依存せずに情報を獲得するために，獲得したデータベースを統合する機能

本稿では，記事などに含まれるドキュメントから複数の情報源にアクセスし，獲得したデータを統合的に並置しデータを横断的に獲得することにより，情報を獲得するための異種のデータベースを統合した閲覧環境を示し，記事に含まれるドキュメントと，法令や書籍といった専門データベースを対象としたデータベース統合システムの実現について述べ，実験によりその実現可能性を示す．

2.　本方式の概要

特定の専門分野を扱っているデータベースを連結し，統合的な閲覧環境を実現する方式の機能群について示す．本方式はニュース記事などのドキュメントを参照する際において，特定の専門分野を対象としているデータベースを対象として閲覧中のドキュメントと関連の深い専門データベース中から自動的に関連の深いコンテンツを検索し，連結して提示することによって統合的な閲覧環境を実現する．(図 1 を参照 2.)

[*機能 *1:*類似度計算に用いるデータベースコンテンツのメタデータ抽出*] 閲覧対象の記事などのドキュメントを 専門データベース群のコンテンツと関連づける際に用いるメタデータを抽出する．メタデータ抽出機能は以下の式によって表される．

$$f_{conversion}(D_i, N) \rightarrow M_i \tag{1}$$

D_i:対象ドキュメント i
i:ドキュメント識別子，$i = 1 \ldots n$
M_i:メタデータ集合 i
N:ドキュメントの特徴付けに必要な名詞のみを抽出した特徴語集合

複数のカテゴリと関連付けられた専門データベース群内の各ドキュメントを形態素解析によって単語に分割する．特徴語集合 N を用いて，ドキュメントの特徴付けに不要な助詞などの単語を除き，出現頻度の上位の単語群をメタデータとして抜き出す．

機能 *2:*和演算機能　本方式では，専門データベース群を分野ごとに分類し，閲覧対象ドキュメントの性質に応じて，閲覧対象ドキュメントと関連を持っているデータベース群のみを結合する．本機能はこのデータベース結合に相当する．

$$f_{union}(S_1^{in}, S_2^{in}) \rightarrow S^{out} \tag{2}$$

S_1^{in}:対象ドキュメント D_i の集合
S_2^{in}:対象ドキュメント D_i の集合
S^{out}:結果ドキュメント集合

閲覧対象ドキュメントを，該当ドキュメントの情報源である記事データベース上に配置し，ドキュメントをカテゴリに分類する．記事データベースにある,「経済」「政治」といった各ジャンルに含まれている記事群を専門データベースに関連する各カテゴリと関連付ける．その閲覧対象ドキュメントのカテゴリと専門データベースのカテゴリ群間において一致検索し，一致する専門データベースのみをアクセスし結合を行う．統合の対象するデータベースを閲覧対象ドキュメントの性質に応じて動的に選択することで，不適切なデータの結合を抑制する．

機能 *3:*閲覧対象ドキュメントと専門分野データベースの統合機能 *(*ドキュメントの類似度計算機能*)*

問い合わせとなる閲覧対象ドキュメントと，選定しアクセスした専門データベース群内のデータとの類似度を計量し，類似度の高いデータ間を統合して，利用者に提示する．本機能に

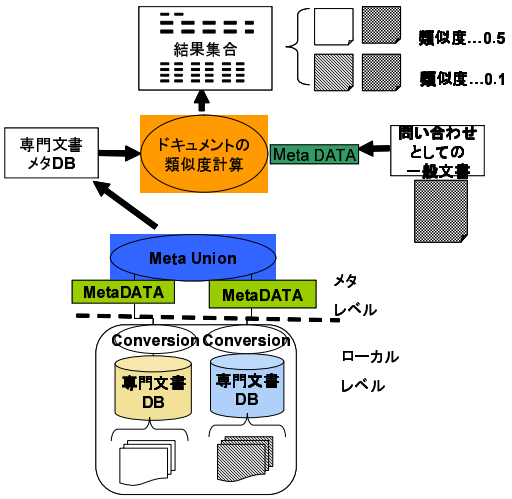

図 1　本方式の概要

より閲覧対象ドキュメントに関連の強い専門情報を自動的に取得可能な環境が実現される．本方式では，機能 1 により生成されたメタデータから閲覧対象ドキュメントデータベース内における各ドキュメントを検索し，該当するドキュメントと検索を行ったメタデータを持つ専門データベースのドキュメント群を並置する．メタデータとドキュメントの類似度の計算は各ドキュメントに含まれるメタデータの頻出度を利用し，以下の方法により行われる．

$$f_{join}(S_1^{in}, S_2^{in}, \alpha) \rightarrow R \qquad (3)$$

α:結合条件.

R:ドキュメントの対，およびそれらの間の相関量の集合.

閲覧対象ドキュメント中に出現する単語群に，専門データベース内のドキュメントが持つメタデータと一致する単語の数を用いて類似度を計量する．結合条件についての詳細は 3 節で述べる．

複数の異種の専門データベース群を統合し，統合的な情報を獲得するために，本方式は，以上の 4 機能を以下の手順で適用する．本方式では，マルチデータベースにおける異種データベース群の統合方式 [1], [3], [5] を用いる．

Step-1　:各専門データベースのコンテンツのメタデータを抽出

Step-2　:問い合わせとなる一般ドキュメントのメタデータ生成

Step-3　:結合されるデータベース群の選択

Step-4　:一般ドキュメントと専門ドキュメント間との類似度計算

3.　本方式の実現

本節では，本方式の実現の詳細について示す.

3.1　相関量計量に用いる専門データベースコンテンツのメタデータ抽出機能

専門データベース群のコンテンツと閲覧対象ドキュメントとする記事などのドキュメントを関連づける際に用いるメタデータを抽出する.

各データベースにおけるドキュメントを関連づけるために，異種の専門データベースを統合するという前提に基づくメタデータとして各専門データベース内において出現する単語をその単語と出現頻度の組として生成する．これにより，各専門データベースの特徴を表現するのに必要な単語群がメタデータ群として蓄積される．メタデータ変換方法は各項目に分類されたデータを形態素解析して行われる．形態素解析には様々なアルゴリズムおよび実装が存在するが，本方式においては形態素解析器として茶筌 [8] を利用した．形態素解析によって得た単語群は，メタデータとして用いる名詞のみを抽出する．各単語の出現頻度の順に並べる．生成された特徴単語データを専門データベース内におけるそれぞれのデータ群のメタデータとして登録する.

3.2　MetaUnion 機能の適用

本実現では，ORDB の有する和演算を用いる．WWW 上の異種のデータベースからの情報を実現するために，本方式ではマルチデータベースシステムである ADMIS [1] を用いる．ADMIS は，既存の WWW 上の情報資源から独立した，一段階抽象度の高いメタレベルシステムである．ADMIS の特徴は，異種の情報資源を統合するために，メタレベルシステム上にデータを動的に写像し，それらの情報の結合条件の判定を動的に算出する点にある．本実現では各機能群を ADMIS のこの動的な写像 $f_{conversion}(D_i, N)$，および，動的な結合条件 α として実装する.

記事データベース上に配置されたドキュメントを動的に獲得するために，記事データベースが持つ特徴を利用して各ドキュメントデータをカテゴリに分類する．記事データベースにある「経済」「政治」といった各ジャンルに含まれている記事群を専門データベースに関連する各カテゴリと関連付け，専門データベースのカテゴリ群間において一致検索し，法律などといった一致する専門データベースのみをアクセスし，データを獲得する.

3.3　閲覧対象ドキュメントと専門分野データベースの統合機能

各専門データベースと閲覧対象ドキュメントの間の類似度を計算するために，本方式では以下の方法を適用する．ここでは，2 節で示した f_{join} における結合条件 α の具体的実現として，g_{doc} および，g_{Auth} の 2 関数を設定する.

(1)　専門語使用率による関数

専門データベース内におけるデータを対象としたドキュメントデータベース内に含まれる各ドキュメントの特徴を g_{doc} により定義する．g_{doc} により出力される値をドキュメントデータベース内における単語群の専門語使用率とする.

$$g_{doc}(S_n, D_m) \rightarrow L_{[n,m]} \qquad (4)$$

S_n:専門データベース内における出現単語の中でドキュメントデータベース内における出現単語と一致する単語群

D_m:ドキュメントデータベース内における出現単語群

$L_{[n,m]}$:ドキュメントデータベース内における単語群内に持つ専門データベースと一致する単語の相関量

(2)　専門語含有率による関数

一般文書データベース内に含まれる各ドキュメントを対象とし

た専門データベース内におけるデータの相関量計算を，g_{Auth} により定義する．g_{Auth} により出力される値を一般文書ドキュメントデータベース内における単語群専門語含有率とする．

$$g_{Auth}(S_n, D_m) \rightarrow A_{[n,m]} \quad (5)$$

S_n:専門データベース内における出現単語群

D_m:ドキュメントデータベース内における出現単語群の中で専門データベース内における出現単語と一致する単語群

$A[n, m]$:ドキュメントデータベース内における単語群内に持つ専門データベースと一致する単語の相関量

4. 実　　験

本節では，報道記事や法令などの異種のデータベース群を対象とした統合的な参照および閲覧を実現する本方式を，実現可能性の観点から評価する．

4.1 実験の目的

実験の目的は以下の 3 項である．

(1) 異種のデータベース群の有する専門データベースおよび，閲覧対象ドキュメントをマルチデータベース環境において結合し，統合的閲覧環境を実現する．

(2) 本方式で用いたドキュメントデータベース内における単語群の専門語使用率 (3.3 節) による計算結果から，関数 g_{doc} の性能について評価する．

(3) 本方式で用いた閲覧対象ドキュメント内における単語群の専門語含有率 (3.3 節) による計算結果を関数 g_{Auth} の性能について評価する．

4.2 実験環境および実験手法

実験では閲覧対象ドキュメントとして WWW 上の Nikkei NET [9] から得たドキュメント 10 件を用いた．

専門データベースとしては，法令データ提供システムから獲得した次の 3 データベースを用いた。

- 公職選挙法データベース: 政治分野に対応するデータベースコンテンツ（10 件）
- 少子化社会対策会議令データベース: 経済分野に対応するデータベースコンテンツ（10 件）
- 介護保険法データベース: 医療分野に対応するデータベースコンテンツ（10 件）

4.3 実 験 結 果

3.1 節で示した相関量計量に用いるデータベースコンテンツのメタデータ抽出機能の実現システムの実行結果を表 1 に示す．3.2 節で示した MetaUnion 機能の実現システムの実行結果を表 2 に示す．3.3 節で示した閲覧対象ドキュメントと専門分野データベースの統合機能における g_{doc} により出力された専門語使用率の実行結果を表 3 に示す．同様に閲覧対象ドキュメントと専門分野データベースの統合機能における g_{Auth} により出力された専門語含有率の実行結果を表 4 に示す．

4.4 考　　察

第 1 に，新聞記事を有する一般的なドキュメントデータベース，および，法令に関する専門的なドキュメントデータベース群を対象として，本方式の有する機能群を適用することで，これらのデータベース群を動的に連結し，異種データベースがそれぞれ有する関連ドキュメントの対を閲覧可能であることを確認した (表 3，表 4)．

本実験によりドキュメントデータベース内における単語群の専門語使用率に出力される g_{doc} の実現可能性を示した．表 3 では，専門データベース中のドキュメントと，それに関連の深い閲覧する記事ドキュメントの対が上位に検索されている．これは，法令などの専門データベースの有するドキュメント群が，一般ドキュメントとの対比において，多数のメタデータを有するため，関連の高い一般ドキュメントの単語を内部的に有することに起因する．

表 4 では，閲覧する記事ドキュメントと，それに関連の高い専門データベースの閲覧対象ドキュメントとの対が上位に検索されなかった．これは，新聞記事のメタデータの語数が，専門ドキュメントとの対比において，少数のメタデータを有し，また，そのメタデータの語群における専門的単語の比率が低いことに起因すると考えられる．

本実験で示した実験データを対象とした比較を行った結果，式 4 で示した専門語使用率が，専門語含有率に比べ，適切な結果を示した．本実現により記事の特徴を反映して関連する専門データベースの統合的な閲覧が可能になった．

以上の実験結果により，閲覧対象ドキュメント群および法令などの異種のデータベース群をマルチデータベース環境において結合する，統合的閲覧環境の実現可能性が示された．

5. ま と め

本稿では，複数の異種データベース群を対象として統合的に参照および閲覧可能な環境の実現方式を示した．

本方式により，異種データベースをマルチデータベース環境に連結・統合を行うことにより，利用者は異種の情報源間に存在する関連情報を動的に参照し，統合的な閲覧を行うことが可能となった．本方式により，あらかじめ対象ドキュメント群と，結合されるデータベース間の関連度を計量し，参照するデータベースを選択することにより，不適切なデータの結合を抑制することが可能となった．

今後の課題として，本方式の定量的および解析的実験，大規模データベースを対象とした実験，実際の利用環境における実現が挙げられる．

謝　　辞

本研究の実現にあたって，多くの御助言を頂いた慶應義塾大学大学院政策・メディア研究科吉田尚史氏，慶應義塾大学大学院政策・メディア研究科石橋直樹氏に感謝致します．

表 1　専門データベースから抽出されたメタデータの総数

法令タイトル	含有率
公職選挙法	61891
少子化社会対策会議令	974
介護保険法	23188

表 2　閲覧対象ドキュメントから抽出されたメタデータの総数

記事タイトル	含有率
学力向上路線を追認、学習指導要領を一部改訂	82
民主、初当選議員に国会ノウハウを集中指導	117
介護施設入所者、住居費負担へ	114

表 3　実験結果 1
(閲覧対象ドキュメントと専門分野データベースの統合機能における g_{doc} により出力された専門語使用率を用いた実行結果)

閲覧対象ドキュメント名 → 関連する記事ドキュメント名	専門語使用率
道路民営化会社株の政府保有比率、３分の１以上に → 少子化社会対策会議令	0.039474
道路民営化会社株の政府保有比率、３分の１以上に → 平成１４年度の経済見通しと経済財政運営の基本的態度	0.025287
道路民営化会社株の政府保有比率、３分の１以上に → 月例経済報告	0.017271
道路民営化会社株の政府保有比率、３分の１以上に → 老人保健法施行規則第十八条の規定に基づき厚生労働大臣が定める収入の額の算定方法	0.015748
道路民営化会社株の政府保有比率、３分の１以上に → 消費者保護基本法	0.011352
道路民営化会社株の政府保有比率、３分の１以上に → 北朝鮮当局によって拉致された被害者等の支援に関する法律に基づく国民年金の特例に関する省令	0.008475
道路民営化会社株の政府保有比率、３分の１以上に → 平成十五年度における国民年金法による年金の額等の改定の特例に関する法律	0.008372
道路民営化会社株の政府保有比率、３分の１以上に → 平成十五年八月から平成十六年七月までの月分の労働者災害補償保険法の規定による年金たる保険給付又は平成十五年八月一日から平成十六年七月三十一日までの間に支給すべき事由が生じた同法の規定による障害補償一時金等に係る給付基礎日額の算定に用いる厚生労働大臣が定める率	0.008104
道路民営化会社株の政府保有比率、３分の１以上に → 金融商品の販売等に関する法律	0.00678
道路民営化会社株の政府保有比率、３分の１以上に → 医師法	0.00594
道路民営化会社株の政府保有比率、３分の１以上に → インフルエンザに関する特定感染症予防指針	0.005765
道路民営化会社株の政府保有比率、３分の１以上に → 老人医療費の伸びを適正化するための指針	0.00496
道路民営化会社株の政府保有比率、３分の１以上に → 月例経済報告	0.004938
道路民営化会社株の政府保有比率、３分の１以上に → 医業若しくは歯科医業又は病院若しくは診療所に関して広告することができる事項	0.00484
道路民営化会社株の政府保有比率、３分の１以上に → 感染症の予防及び感染症の患者に対する医療に関する法律	0.004808
道路民営化会社株の政府保有比率、３分の１以上に → 金融機能の早期健全化のための緊急措置に関する法律	0.003259
道路民営化会社株の政府保有比率、３分の１以上に → 証券会社に関する内閣府令	0.00155
道路民営化会社株の政府保有比率、３分の１以上に → 介護保険法施行規則	0.001012
道路民営化会社株の政府保有比率、３分の１以上に → 介護保険法	0.000662
道路民営化会社株の政府保有比率、３分の１以上に → 公職選挙法	0.000297

表 4　実験結果 2

(閲覧対象ドキュメントと専門分野データベースの統合機能における g_{Auth} により出力された専門語含有率を用いた実行結果)

閲覧対象ドキュメント名 → 関連する法令名	専門語含有率
民主、初当選議員に国会ノウハウを集中指導 → 介護保険法	0.359649123
介護施設入所者、住居費負担へ → 公職選挙法	0.280487805
民主、初当選議員に国会ノウハウを集中指導 → 公職選挙法	0.245614035
介護施設入所者、住居費負担へ → 介護保険法	0.207317073
民主、初当選議員に国会ノウハウを集中指導 → 少子化社会対策会議令	0.122807018
介護施設入所者、住居費負担へ → 少子化社会対策会議令	0.12195122

文　　献

[1] Shuichi Kurabayashi and Yasushi Kiyoki, ”A Meta-Level Active Multidatabase System Architecture for Heterogeneous Information Resources” Information Modelling and Knowledge Bases (IOS Press), Vol. 15, June 2003.

[2] Y. Kiyoki, T. Kitagawa and T. Hayama, “A metadatabase system for semantic image search by a mathematical model of meaning, ACM SIGMOD Record, (refereed as the invited paper for special issue on metadata for digital media), Vol.23, No. 4, pp.34-41, Dec. 1994.

[3] N. Ishibashi, Y. Hosokawa, and Y. Kiyoki, “A Spatial and Temporal Data Integration Method for Heterogeneous Database Environments”, Proceedings of the 19th IASTED International Conference on APPLIED INFORMATICS (AI2001), pp.323–330, Feb. 2001.

[4] S. Kurabayashi, N. Ishibashi and Y. Kiyoki: “A Multidatabase System Architecture for Integrating Heterogeneous Databases with Meta-Level Active Rule Primitives”, Proceedings of the 20th IASTED International Conference on APPLIED INFORMATICS (AI2002), Feb. 2002.

[5] Y. Kiyoki, Y. Hosokawa and N. Ishibashi: “A Metadatabase System Architecture for Integrating Heterogeneous Databases with Temporal and Spatial Operations,” Advanced Database Research and Development Series Vol. 10, Advances in Multimedia and Databases for the New Century, A Swiss/Japanese Perspective, pp.158–165, World Scientific Publishing, 2000.

[6] 石橋 直樹, 細川 宜秀, 清木 康: “時空間的文脈に応じた動的関連性計量機構を有する異種データベース間結合方式”, 情報処理学会論文誌:データベース, Vol.43, No.SIG2(TOD13), pp.128–145, 2002.

[7] 河本穣, 清木康, 吉田尚史, 藤島清太郎, 相磯貞和, ” 医療分野ドキュメント群を対象とした意味的連想検索空間の実現方式,” 日本データベース学会 Letters, Vol.1, No.2, March 2003, pp. 12-15.

[8] 日本語形態素解析システム『茶筌』http://chasen.aist-nara.ac.jp/

[9] NIKKEI NET http://www.nikkei.co.jp/