研究会プログラムのWebページからの抽出

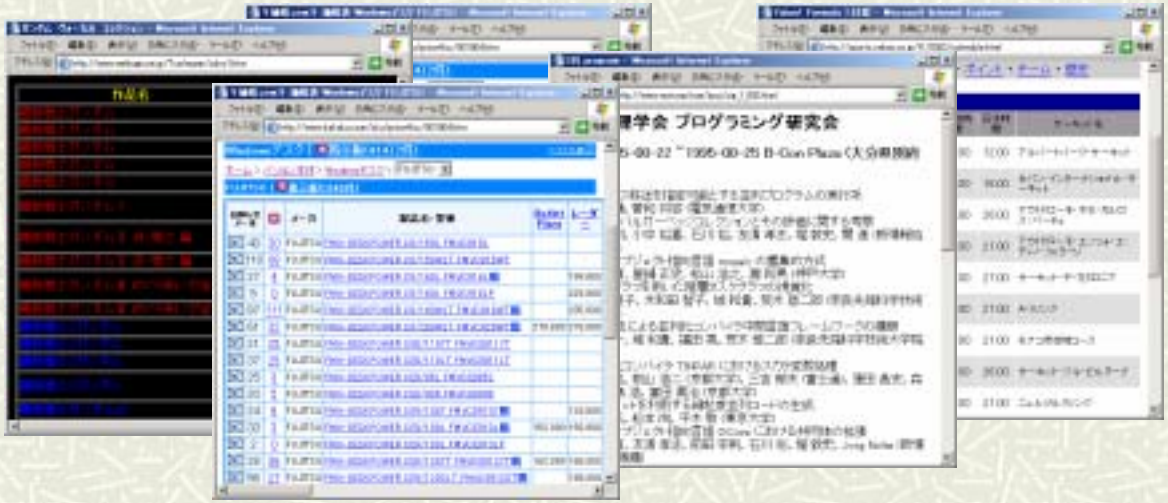
渡辺精一郎 九州大学システム情報科学府
廣川佐千男 九州大学情報基盤センター

DEWS2002 2002/3/4

背景

- Web文書（ページ）：半構造化データ
  - いくつかのデータから成る組が繰り返し出現するページが多数存在・・・表、名簿をもつページなど

レコード



各ページからのレコードの抽出 → Web上の情報をデータベースの様に扱うことが可能

Webページ中のレコード

- 各ページによって表現や形式が異なる
  - 表現形態：表、リスト、箇条書き
  - HTML、XMLのタグの利用

簡単にレコード抽出はできない

- 一つのページに注目：レコードの表現や形式は同じ

表現、形式に関する情報からレコードの抽出が可能

抽出対象

- 対象：研究会プログラムが記載されたページ
  - 特徴：テキスト(タグを用いていない)ページが多い
- 抽出するもの
  - ページ中の各レコード
  - 各レコードに共通する情報

COMMONDATA

レコード、COMMONDATAに関する情報をページのフォーマット情報とする

レコードとCOMMONDATA

SIG program - Microsoft Internet Explorer

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H) リンク

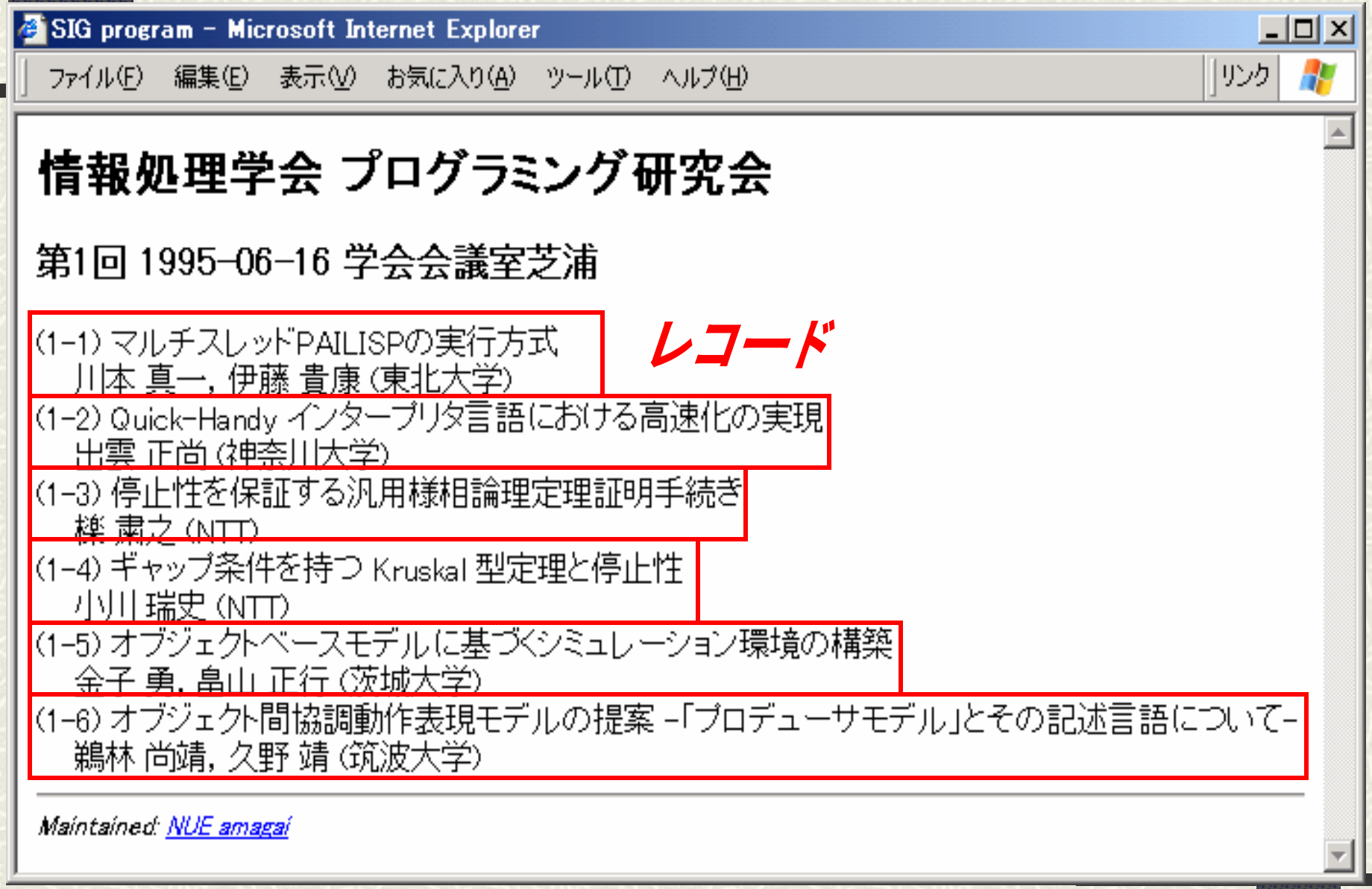
情報処理学会 プログラミング研究会

第1回 1995-06-16 学会会議室芝浦

(1-1) マルチスレッドPAILISPの実行方式
　川本 真一, 伊藤 貴康 (東北大学)
(1-2) Quick-Handy インタープリタ言語における高速化の実現
　出雲 正尚 (神奈川大学)
(1-3) 停止性を保証する汎用様相論理定理証明手続き
　楳 粛之 (NTT)
(1-4) ギャップ条件を持つ Kruskal 型定理と停止性
　小川 瑞史 (NTT)
(1-5) オブジェクトベースモデルに基づくシミュレーション環境の構築
　金子 勇, 畠山 正行 (茨城大学)
(1-6) オブジェクト間協調動作表現モデルの提案 -「プロデューサモデル」とその記述言語について-
　鵜林 尚靖, 久野 靖 (筑波大学)

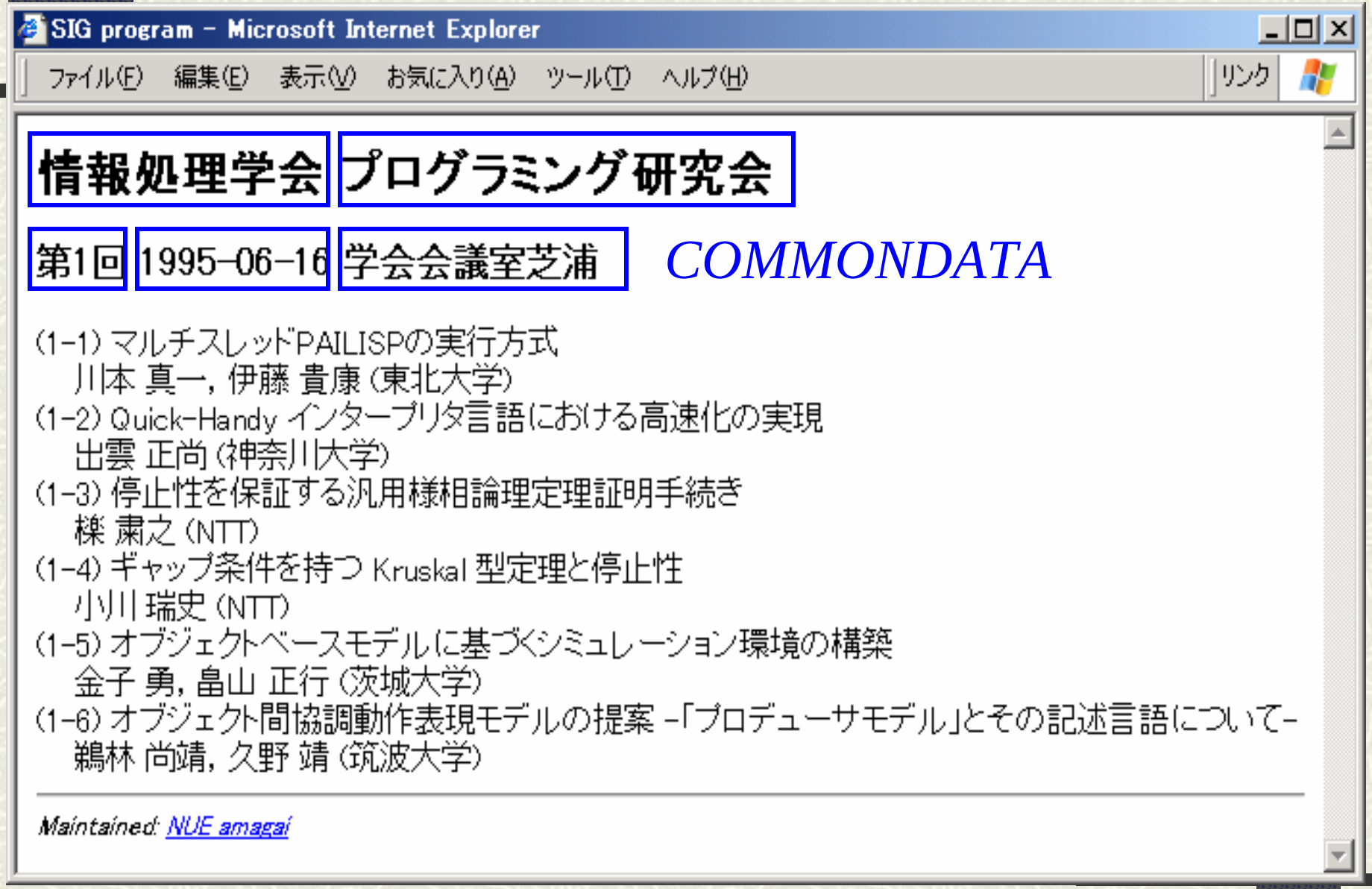
Maintained: NUE amagai

レコードとCOMMONDATA



レコード

レコードとCOMMONDATA



COMMONDATA

- 以下の4つの情報で表記
  - HEAD
    - 抽出する対象の直前の文字列
  - TAIL
    - 抽出する対象の直後の文字列
  - H-NUM(HEAD-NUMBER)
    - HEADが文書中で何番目なのかを表す
  - T-NUM(TAIL-NUMBER)
    - TAILがHEAD以降で何番目なのか表す

レコード情報

- レコード群の出現領域を指定
  - R－HEAD
    - 領域の直前の文字列
  - R－TAIL
    - 領域の直後の文字列
  - RH－NUM
    - R-HEADが文書中で何番目なのかを表す
  - RT－NUM
    - R-TAILがHEAD以降で何番目なのか表す

レコード情報

- レコードの形式に関する情報
  - DELIMITER
    - 1つのレコード中の各フィールド間を区切っている文字情報
  - ITEM
    - 各フィールドの項目名

この2つをフィールドの数に応じて記述

レコード抽出プログラム

- 一つのページからページ中の各レコード、COMMONDATAを抽出
- 入力
  - ページのソース
  - ページのフォーマット情報
- 出力
  - レコード
  - COMMONDATA

レコード抽出実験

対象：情報処理学会所属の8研究会のプログラムを記載したページ（計133ページ）

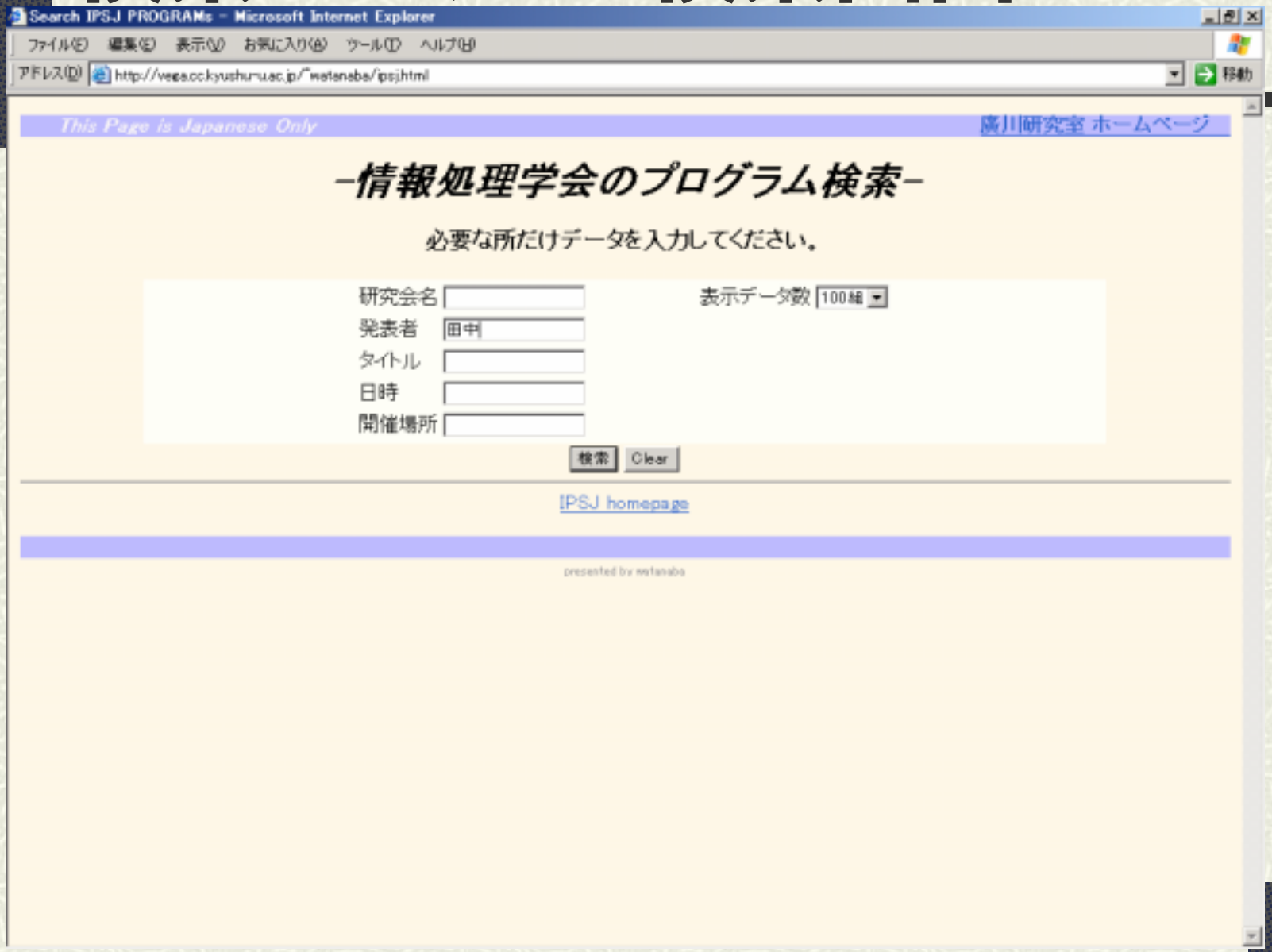
- タイトル、発表者の2つの情報は必ず抽出
- 他に以下の情報がページ中にあれば抽出

（日時、開催場所、主査、幹事、議題、学会名、研究会名、開催回数、発表時間、要約）

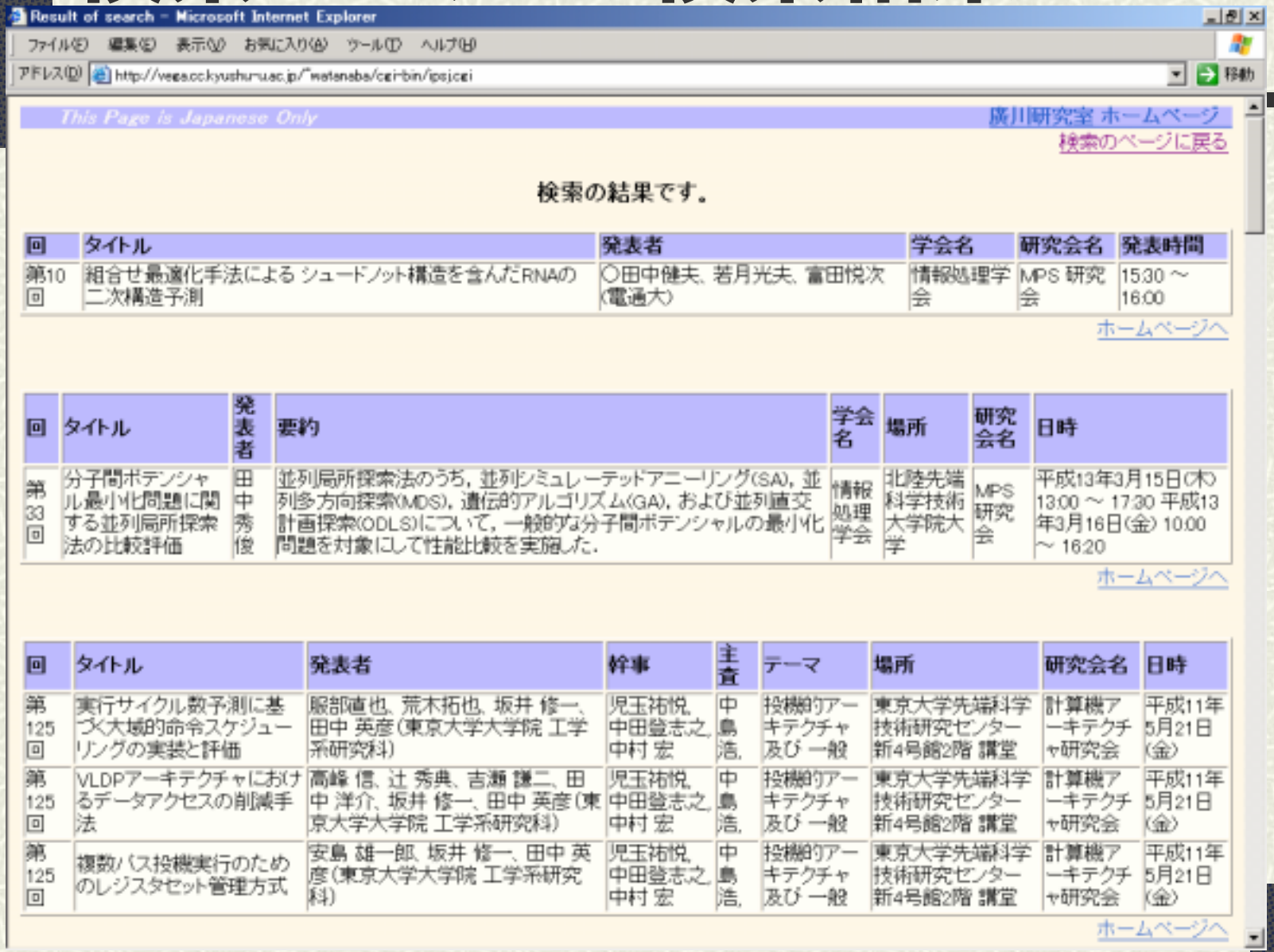
レコード抽出実験結果

研究会名	ページ数	成功(完全)	失敗	定義不可
DBS	11	4(4)	1	6
ARC	9	1(0)	0	8
SLDM	9	7(7)	0	2
PRO	32	32(32)	0	0
AL	16	10(0)	0	6
MPS	33	19(19)	14	0
HPC	17	16(11)	0	1
OS	6	0(0)	0	6
計	**133**	**89(73)**	**15**	**29**

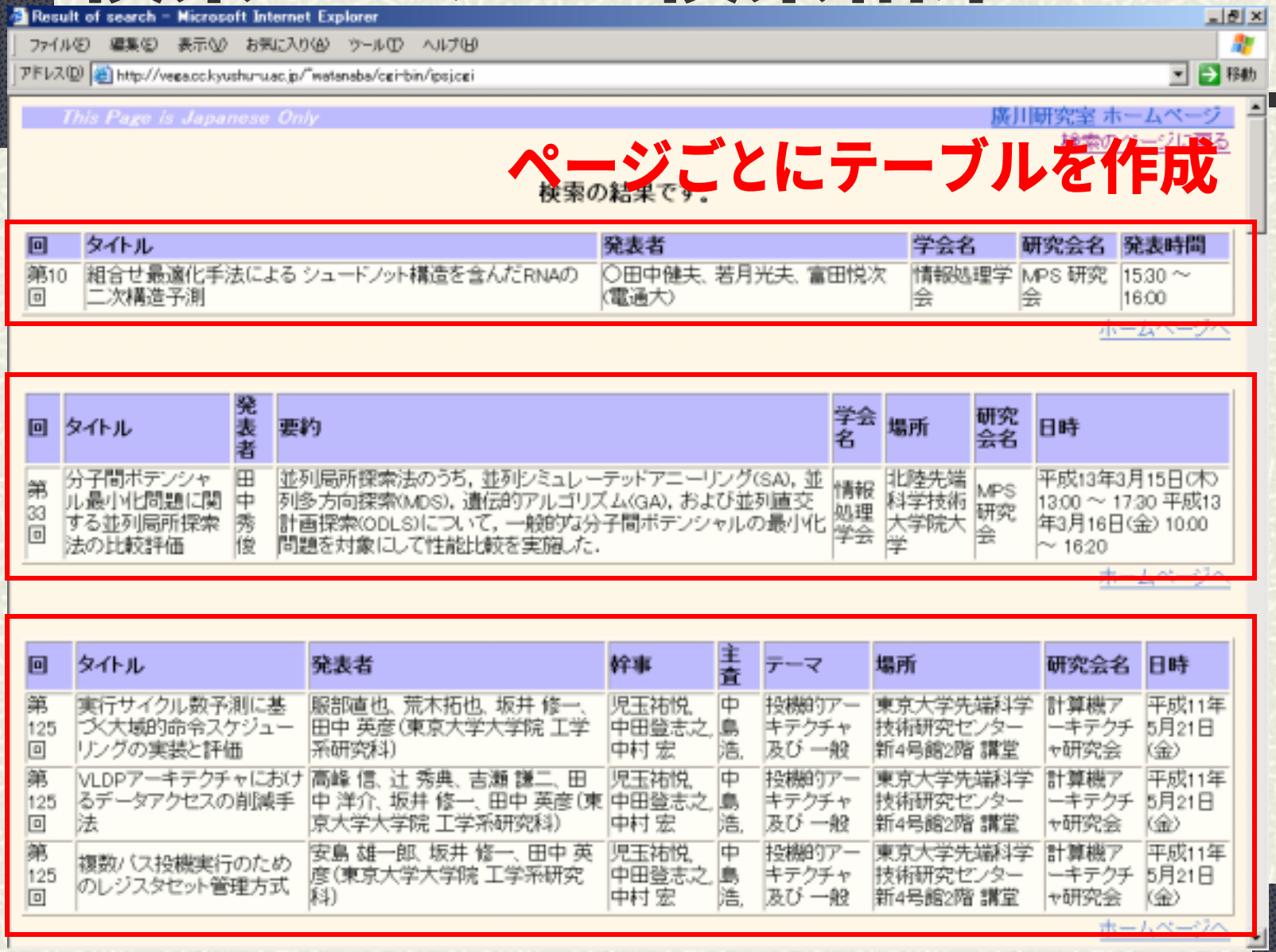
検索システム：検索画面


Search IPSJ PROGRAMs - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://vega.cc.kyushu-u.ac.jp/~watanaba/ipsj.html
移動
This Page is Japanese Only
廣川研究室 ホームページ
-情報処理学会のプログラム検索-
必要な所だけデータを入力してください。
研究会名
発表者
田中
タイトル
日時
開催場所
表示データ数
100組
検索
Clear
IPSJ homepage
presented by watanaba

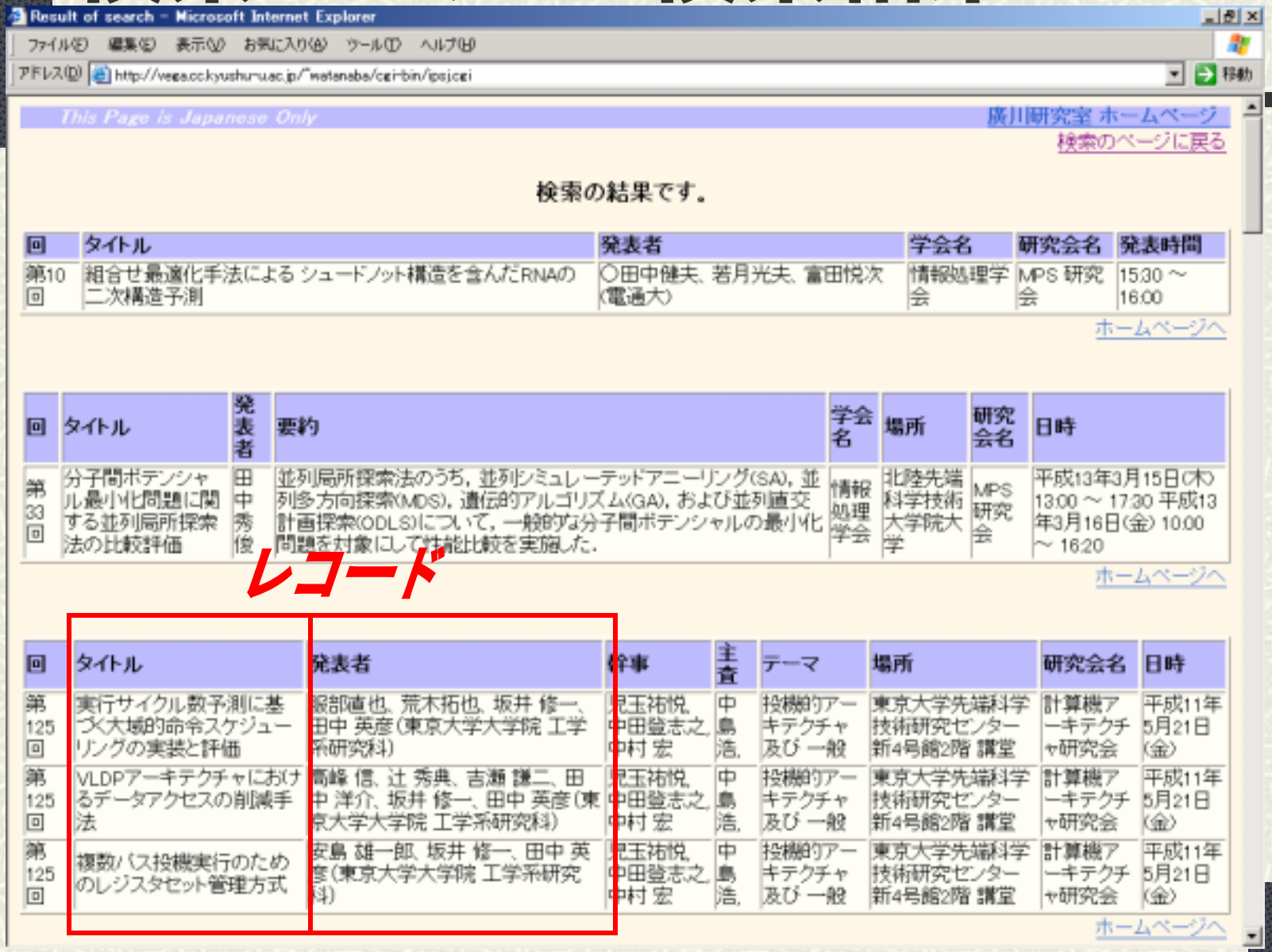
検索システム：検索結果


Result of search - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://vega.cc.kyushu-u.ac.jp/~watanaba/cgi-bin/ipsj.cgi
移動
This Page is Japanese Only
廣川研究室 ホームページ
検索のページに戻る
検索の結果です。
回 タイトル 発表者 学会名 研究会名 発表時間
第10回 組合せ最適化手法による シュードノット構造を含んだRNAの二次構造予測 ○田中健夫、若月光夫、富田悦次(電通大) 情報処理学会 MPS 研究会 15:30 ～ 16:00
ホームページへ
回 タイトル 発表者 要約 学会名 場所 研究会名 日時
第33回 分子間ポテンシャル最小化問題に関する並列局所探索法の比較評価 田中秀俊 並列局所探索法のうち，並列シミュレーテッドアニーリング(SA)，並列多方向探索(MDS)，遺伝的アルゴリズム(GA)，および並列直交計画探索(ODLS)について，一般的な分子間ポテンシャルの最小化問題を対象にして性能比較を実施した． 情報処理学会 北陸先端科学技術大学院大学 MPS研究会 平成13年3月15日(木) 13:00 ～ 17:30 平成13年3月16日(金) 10:00 ～ 16:20
ホームページへ
回 タイトル 発表者 幹事 主査 テーマ 場所 研究会名 日時
第125回 実行サイクル数予測に基づく大域的命令スケジューリングの実装と評価 服部直也、荒木拓也、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び 一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
第125回 VLDPアーキテクチャにおけるデータアクセスの削減手法 高峰 信、辻 秀典、吉瀬 謙二、田中 洋介、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び 一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
第125回 複数パス投機実行のためのレジスタセット管理方式 安島 雄一郎、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び 一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
ホームページへ

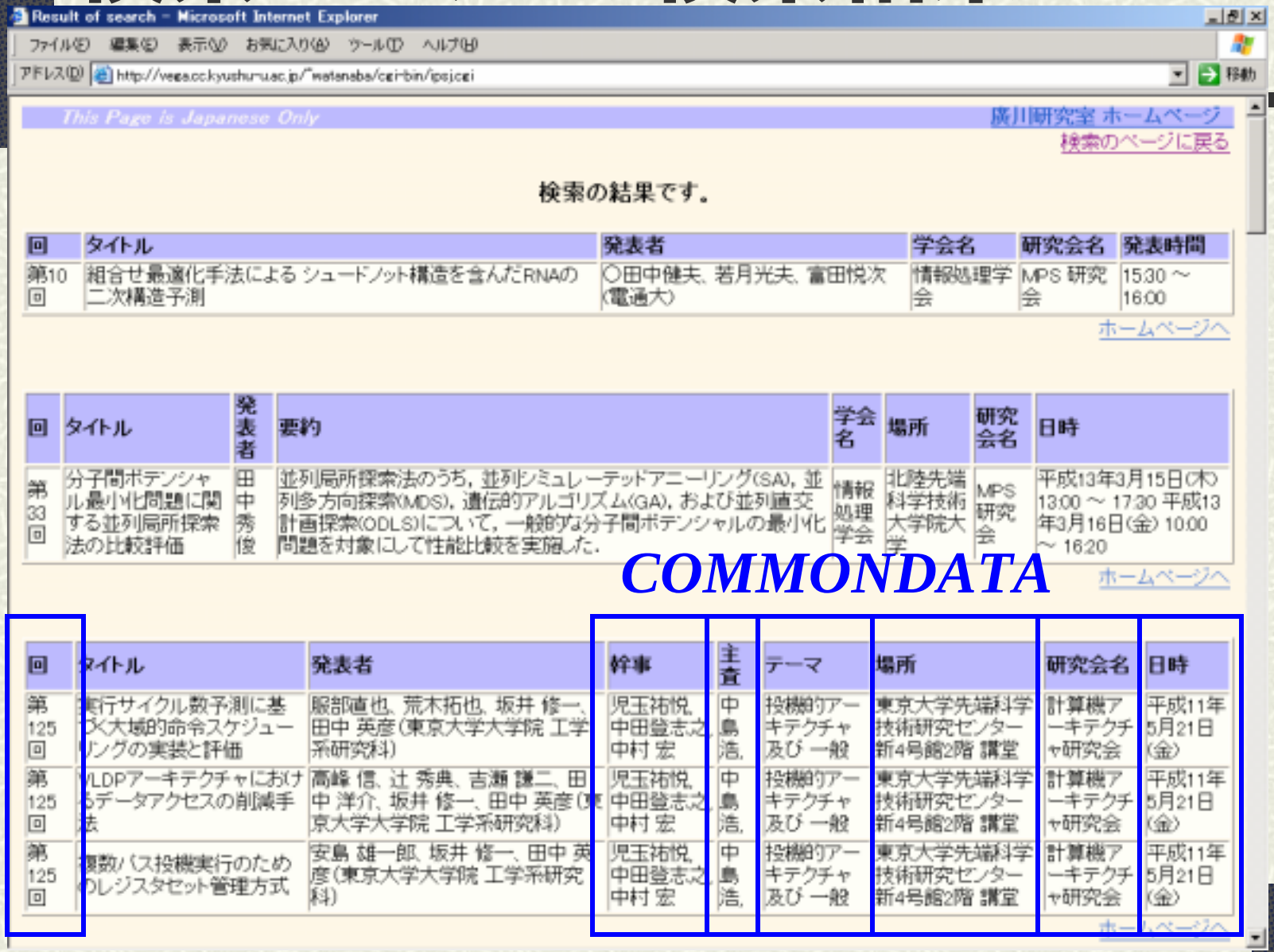
検索システム：検索結果


Result of search - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://vega.cc.kyushu-u.ac.jp/~watanaba/cgi-bin/ipsj.cgi
移動
This Page is Japanese Only
廣川研究室 ホームページ
検索のページに戻る
検索の結果です。
ページごとにテーブルを作成
回 タイトル 発表者 学会名 研究会名 発表時間
第10回 組合せ最適化手法による シュードノット構造を含んだRNAの二次構造予測 ○田中健夫、若月光夫、富田悦次(電通大) 情報処理学会 MPS 研究会 15:30 ～ 16:00
ホームページへ
回 タイトル 発表者 要約 学会名 場所 研究会名 日時
第33回 分子間ポテンシャル最小化問題に関する並列局所探索法の比較評価 田中秀俊 並列局所探索法のうち，並列シミュレーテッドアニーリング(SA)，並列多方向探索(MDS)，遺伝的アルゴリズム(GA)，および並列直交計画探索(ODLS)について，一般的な分子間ポテンシャルの最小化問題を対象にして性能比較を実施した． 情報処理学会 北陸先端科学技術大学院大学 MPS研究会 平成13年3月15日(木) 13:00 ～ 17:30 平成13年3月16日(金) 10:00 ～ 16:20
ホームページへ
回 タイトル 発表者 幹事 主査 テーマ 場所 研究会名 日時
第125回 実行サイクル数予測に基づく大域的命令スケジューリングの実装と評価 服部直也、荒木拓也、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
第125回 VLDPアーキテクチャにおけるデータアクセスの削減手法 高峰 信、辻 秀典、吉瀬 謙二、田中 洋介、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
第125回 複数パス投機実行のためのレジスタセット管理方式 安島 雄一郎、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
ホームページへ

検索システム：検索結果


Result of search - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://vega.cc.kyushu-u.ac.jp/~watanaba/cgi-bin/ipsj.cgi
移動
This Page is Japanese Only
廣川研究室 ホームページ
検索のページに戻る
検索の結果です。
回 タイトル 発表者 学会名 研究会名 発表時間
第10回 組合せ最適化手法による シュードノット構造を含んだRNAの二次構造予測 ○田中健夫、若月光夫、富田悦次(電通大) 情報処理学会 MPS 研究会 15:30 ～ 16:00
ホームページへ
回 タイトル 発表者 要約 学会名 場所 研究会名 日時
第33回 分子間ポテンシャル最小化問題に関する並列局所探索法の比較評価 田中秀俊 並列局所探索法のうち，並列シミュレーテッドアニーリング(SA)，並列多方向探索(MDS)，遺伝的アルゴリズム(GA)，および並列直交計画探索(ODLS)について，一般的な分子間ポテンシャルの最小化問題を対象にして性能比較を実施した． 情報処理学会 北陸先端科学技術大学院大学 MPS研究会 平成13年3月15日(木) 13:00 ～ 17:30 平成13年3月16日(金) 10:00 ～ 16:20
ホームページへ
レコード
回 タイトル 発表者 幹事 主査 テーマ 場所 研究会名 日時
第125回 実行サイクル数予測に基づく大域的命令スケジューリングの実装と評価 服部直也、荒木拓也、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
第125回 VLDPアーキテクチャにおけるデータアクセスの削減手法 高峰 信、辻 秀典、吉瀬 謙二、田中 洋介、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
第125回 複数パス投機実行のためのレジスタセット管理方式 安島 雄一郎、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
ホームページへ

検索システム：検索結果


Result of search - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://vega.cc.kyushu-u.ac.jp/~watanaba/cgi-bin/ipsj.cgi
移動
This Page is Japanese Only
廣川研究室 ホームページ
検索のページに戻る
検索の結果です。
回 タイトル 発表者 学会名 研究会名 発表時間
第10回 組合せ最適化手法による シュードノット構造を含んだRNAの二次構造予測 ○田中健夫、若月光夫、富田悦次(電通大) 情報処理学会 MPS 研究会 15:30 ～ 16:00
ホームページへ
回 タイトル 発表者 要約 学会名 場所 研究会名 日時
第33回 分子間ポテンシャル最小化問題に関する並列局所探索法の比較評価 田中秀俊 並列局所探索法のうち，並列シミュレーテッドアニーリング(SA)，並列多方向探索(MDS)，遺伝的アルゴリズム(GA)，および並列直交計画探索(ODLS)について，一般的な分子間ポテンシャルの最小化問題を対象にして性能比較を実施した． 情報処理学会 北陸先端科学技術大学院大学 MPS研究会 平成13年3月15日(木) 13:00 ～ 17:30 平成13年3月16日(金) 10:00 ～ 16:20
ホームページへ
COMMONDATA
回 タイトル 発表者 幹事 主査 テーマ 場所 研究会名 日時
第125回 実行サイクル数予測に基づく大域的命令スケジューリングの実装と評価 服部直也、荒木拓也、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
第125回 VLDPアーキテクチャにおけるデータアクセスの削減手法 高峰 信、辻 秀典、吉瀬 謙二、田中 洋介、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
第125回 複数パス投機実行のためのレジスタセット管理方式 安島 雄一郎、坂井 修一、田中 英彦(東京大学大学院 工学系研究科) 児玉祐悦、中田登志之、中村 宏 中島 浩、 投機的アーキテクチャ及び一般 東京大学先端科学技術研究センター新4号館2階 講堂 計算機アーキテクチャ研究会 平成11年5月21日(金)
ホームページへ

まとめ

- ページのフォーマット情報の表記法の提案
- レコード抽出プログラムとその実験、評価
- 研究会プログラム検索システム

今後の課題

- ページのフォーマット情報の自動抽出
  - 関連研究との連携
- レコードの内容特定
  - 文字列の類似度を比較
  - 既に収集したデータとの比較