パネル討論

高速グラフマイニングが切り拓く世界

1

モデレータ： 鬼塚 真（大阪大学）
パネリスト： Jeffrey Xu Yu（香港中文大学）
中澤 仁（慶応大学），新熊 亮一（京都大学）
諏訪 博彦（奈良先端），藤原 靖宏（NTT）

パネリスト

- Jeffrey Xu Yu(香港中文大学)
- 新熊 亮一(京都大学)
- 諏訪 博彦(奈良先端)
- 藤原 靖宏(NTT)




鬼塚研でのグラフの取り組み

- Hoshizora: 並列グラフエンジン
  - 100コアでも高スケールなBSPエンジン
- 属性付きグラフのクラスタリングアルゴリズム
  - グラフ構造と属性をバランスする隠れ変数を学習
- リンク予測(グラフの時系列変化)
  - NMFによる特徴量抽出 + wolt-winters による予測
- その他:自動OLAP分析, NoSQL自動スキーマ設計, 脳波解析, 特許の類似検索

グラフDBの普及

- クラウド環境でもグラフDB
  - Amazon Neptune
  - Gremlin on Azure Cosmos DB
  - Oracle Spatial and Graph
- 標準化の動向
  - LDBC
  - SQL 標準でのグラフ機能

グラフの分析パターン

- 影響力のある人の検出: PageRank
- パターンマッチング: graph query
- 推薦： NMF, Personalized PageRank
- kNNグラフ＋高速多次元検索
- 半構造データ処理：
- 可視化：

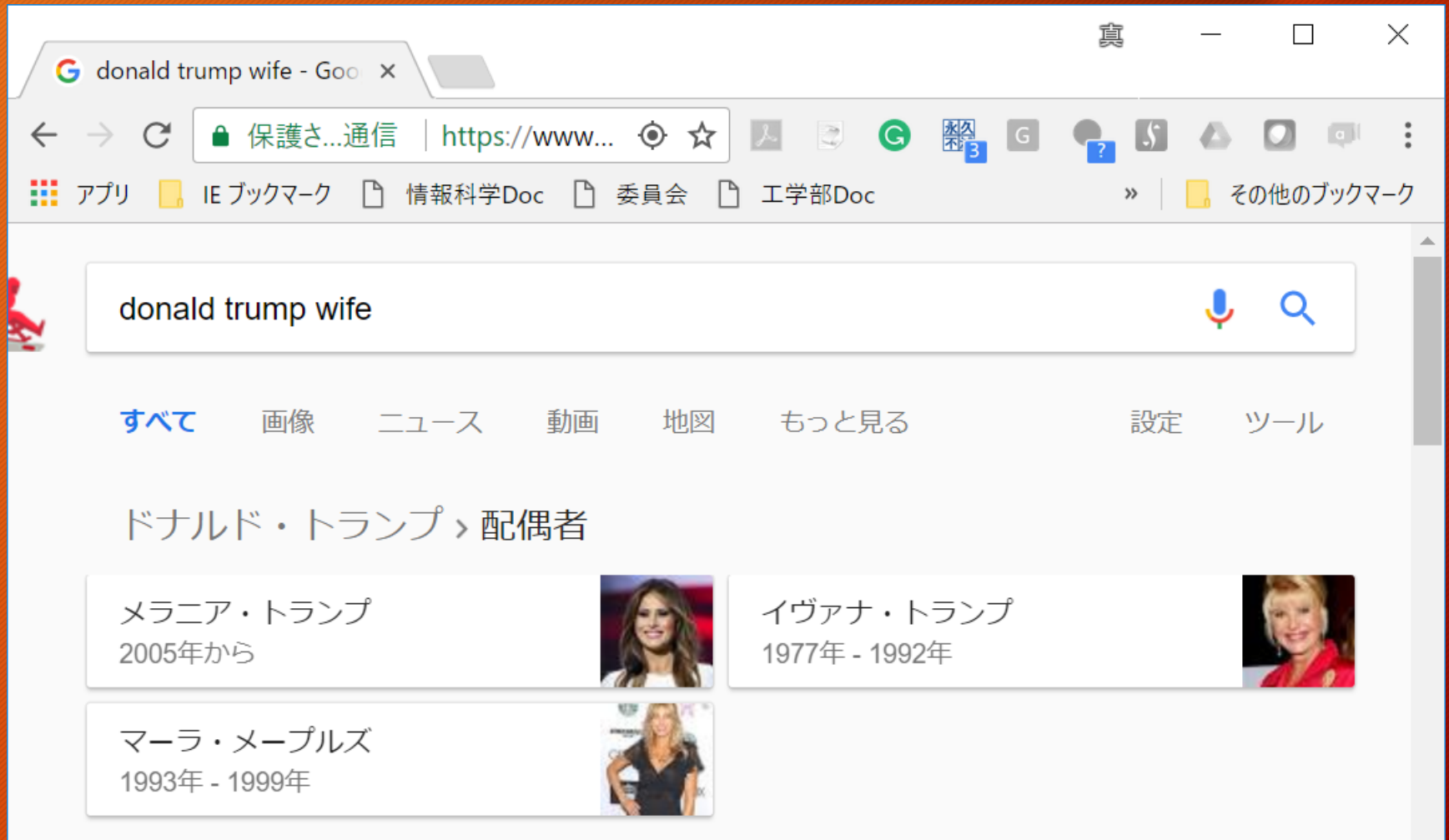
グラフのデータ

- 影響力のある人の検出: ソーシャルグラフ
- パターンマッチング: 化学式
- 推薦：人とモノの関係（購買履歴）
- kNNグラフ：画像，文書（特許，論文）
- 半構造データ処理：ナレッジグラフ
- 可視化：論文引用，共著者，共購買

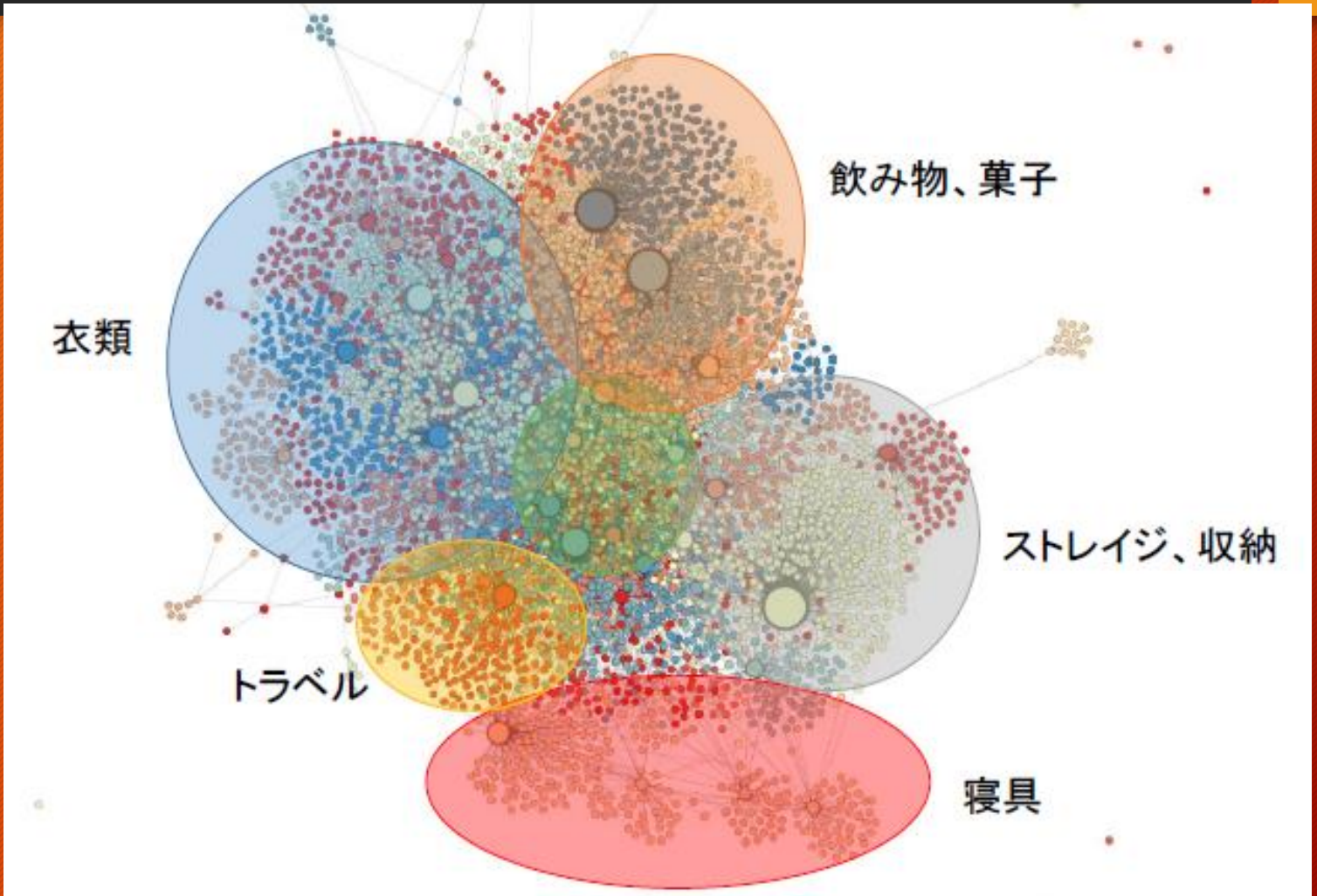
本日の講演でいえば．．．

	Jeffrey	佐藤	諏訪	新熊	榊	荒木	飯田
データ	共著者グラフ(+位置情報)	購買(人×商品)	social graph	動的グラフ	ブログ，twitter，2ちゃん	多種多様	交通網，画像，論文
分析	コミュニティ検知・検索	ロングテール分析	情報伝搬の時系列変化	コンテキストアウェアネス	セグメント特定，情報拡散の分析	Gremlinによる多様な分析	渋滞予想，ラベル推定，分析
技術	高精度化，パラメータ削減	ノイズ除去	次数・媒介中心性	グラフ化	クラスタリング，x2乗	最短ルート・クラスタリング	クラスタリング，ラベル伝搬
応用	理解する	推薦	理解する→予測する	マッチング・可視化	理解する	理解する	予測

ナレッジグラフの利用例



人・モノのグラフ可視化の例



震災ビッグデータの例

出典：http://www.nhk.or.jp/datajournalism/

The Ubiquity of Large Graphs and Surprising Challenges of Graph Processing, VLDB2017

- Neo4j が突出して利用されているわけではない
- Cytoscape は利用者が多い
- 1B edges を超える大規模グラフがかなり多い
- 更に半数以上は dynamic graph だったりする
- 代表的な処理としては Connected Components と clustering が上位2件
- Product-order-transaction グラフが最も多い

The Ubiquity of Large Graphs and Surprising Challenges of Graph Processing, VLDB2017

- 1B edges を超える大規模グラフがかなり多い

Table 6: Sizes of organization that have graphs with >1B edges.

Size	1 - 10	10 - 100	100 - 1000	>10000
#	4	4	7	4

The Ubiquity of Large Graphs and Surprising Challenges of Graph Processing, VLDB2017

- Connected Components と clustering が上位

Computation	Total	R	P	A
Finding Connected Components	55	18	37	12
Neighborhood Queries (e.g., finding 2-degree neighbors of a vertex)	51	19	32	3

(a) Machine learning computations.

Computation	Total	R	P	A
Clustering	42	22	20	15
Classification	28	10	18	2
Regression (Linear / Logistic)	11	5	6	2

パネルの論点

- グラフ処理の研究課題
- グラフを活用した社会貢献
- その他

グラフ処理の研究課題

- 観点
  - グラフの利点を活かした検索支援
  - コミュニケーション支援
  - 推薦，予測
- 実際に，皆さんが実用で困っていることは？
  - グラフDBでどこで使えるの？（適材適所）

グラフを活用した社会貢献

- 観点
  - 機械学習、自然言語処理、センサーデータ分析などの分野と組み合わせて、どんな新サービスを創造できるのか
  - どのように企業活動・行政のスマート化・効率化に貢献できるのか
- 何でビジネス・社会貢献ができるか/そうか？

追加：グラフを使うのメリット

- グラフ構造
  - 自然なモデル化（道路網，ソーシャル，化学式）
  - 多様なデータの活用（多次元とグラフ）
  - 高速化（kNNグラフ）
- グラフ処理
  - グラフ独自の概念：伝搬・経路探索
  - グラフ上で一般的な操作（シミュレーション，距離，クラスタリング）
  - 理解から予測へ

追加：議論の論点

- グラフDB使うの難しい？
  - データのグラフ化が大変
  - 何をノードにして何をエッジにするか？
- データの多様化に伴い分析も複雑化
  - 属性＋グラフの活用
  - 分析精度向上の問題

追加：議論の論点

- グラフクラスタンリングの適用先
- ノイズ削減による大規模データの削減
- SNSでの情報伝搬を捉えて活用する
- ユーザセグメント特定・インサイト発見
- 複数エンジンの最適な組み合わせ

まとめ

- グラフ処理の研究課題
  - Graph processing on RDBMS, network embedding,
  - 動的大規模グラフ, 簡単に使いたい, edge computing
  - グラフ生成, グラフの無限の可能性
  - 構造データ, 非構造データ
- グラフを活用した社会貢献

発表のメモ

Jeffrey 先生

- Community search (local search)
  - OCS: Overwrap community search: no cohesive guarantee, 3 parameters,
  - Querying K-Truss community, SIGMOD 2014
  - The same problem of vertex-cut problem.
  - Question: large cluster is better?
- Local community detection, PVLDB2015
  - Closet community search, PVLDB2016
- Data:
  - Social network, location-based network, transform
- Research topic
  - Dynamic graph

佐藤先生

- レシピと食材の定番度：万願寺とジャコの組み合わせ
- 大規模データの分析が課題
  - ロングテールの問題を解決したい．Z-score を使って，ノイズとノイズではないものを切り分ける．
  - 244万ユーザ×1780万商品の分析
  - データは楽天レビューで，購買順序グラフを作る
  - 周辺部に隠れた名品があるのでは？
- モチーフ分析
  - 質問：購買順序って重要なの？　購買の依存関係がある商品とは何か？
  - コミックのシリーズもの，布団，枕，枕カバー

諏訪先生

- 理解と活用
- 震災時の情報流通を支える
  - 中心性の高いtwitterアカウントの特定
- 楽しむためのツールから情報収集ツールへ：インフラ化
- 震災後では信頼性，専門性の高いアカウントが重要．サザエさんbotのような発信力の強いアカウントの影響がある
- グラフ分析+k-means: クラスタリングして傾向をつかんだことが重要？？
- 中心性＋可視化：

新熊先生

- 関係性技術
  - 可視化・予測：意外性，特異性，共感性
  - 構造から予測ができる（リンク予測）
- Dynamic graph によるコンテキストアウェアネス
- 「おもりんく」：コンテキスト(TPOC)に応じてイベント推薦
- 産業フォーラム：80社
  - いもづる大辞典，関係性オープンツール
  - ワコール：クリック率 254%増（マッチング技術）
  - 川崎重工：CRM分析によるアフターサービスの高度化（可視化）
- 大学発ベンチャー設立：PrediXT
  - 凸版印刷の電子チラシ：マッチング技術
- グラフ分析の適する領域・適さない領域

榊さん

- ホットリンク社の説明
- ユーザ属性推定：セグメンテーション特定（年代，性別）
  - インタラクションを用いてネットワーク化し，クラスタを抽出
  - セグメントを抽出可能
- インサイトの発見
- 情報拡散の分析
  - 実拡散をトラックする研究：伝搬パスが重要なのでは？ Topic rank のような話
- ソフトクラスタリング

荒木さん

- RDBMS, graph engine, full-text search の連携
- 技術：検索の可視化，ファジーな検索
- 最短ルート・クラスタリングに期待
- Azure cosmosDB, Amazon neptune,
  - 推薦，不正検出，知識グラフ，分子構造パターンの探索，
- Apache Tinkerpop:
  - Gremlin で探索が可能
  - OLTP/OLAPの処理が可能
  - インデックスの作り方や性能が影響する
- 自然言語処理