ビッグデータ分析技術ワークショップ
～大規模グラフマイニング技術と応用～

Innovative R&D by NTT

高速グラフマイニング技術 Grapon による大規模グラフ処理と応用

NTTソフトウェアイノベーションセンタ
2017年3月5日

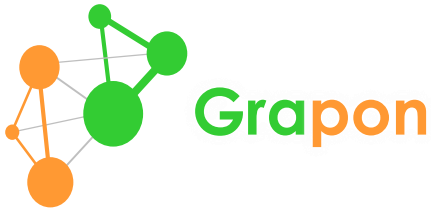
Grapon

本日の発表内容

高速グラフマイニング技術Graponの中から今回，グラフベースの機械学習として有望なラベル伝播を中心に，適用と大規模化対応の研究をご紹介．

＜トピックス＞

- はじめに
- データどうしの類似度によるグラフ構築
- ラベル伝播による文書分類
- 大規模化対応



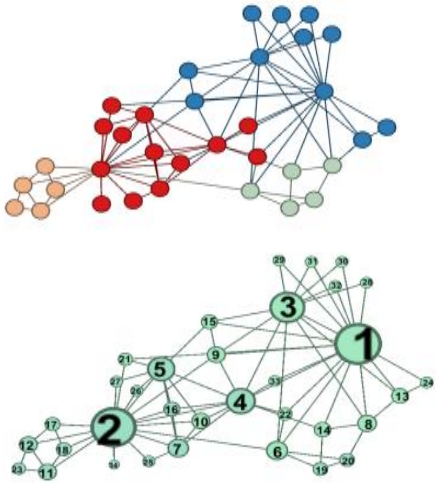
はじめに

- グラフマイニングの利点
- 研究開発の注力分野
- グラフマイニング研究に関わる諸活動

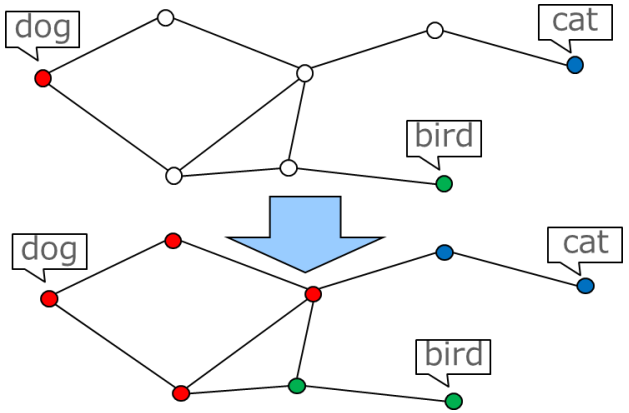
グラフマイニングの利点

- つながり構造を**効率的に処理**可能
- グラフ構造があることで**教師データが少なくても高精度な分析**が可能
- **可視化**との親和性が高い

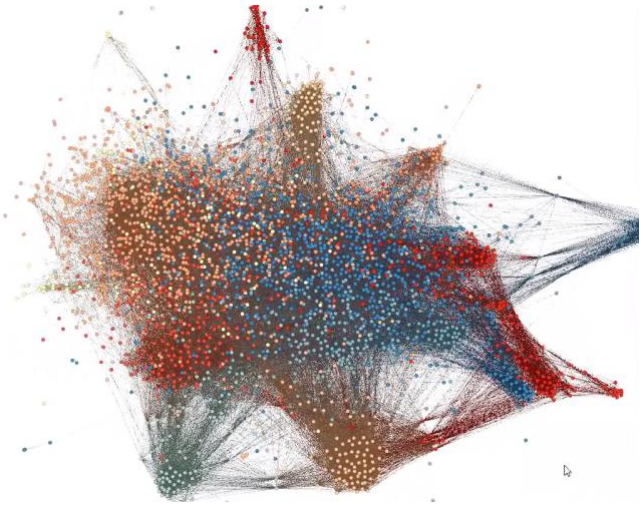
つながりを効率的に処理



少ない教師データから高精度に分析
（ラベル伝播）



可視化による情報把握の容易化
特徴的な部分をすばやく発見



研究開発の注力分野

Innovative R&D by NTT

グラフを活用した分析は有望だが，大規模グラフの分析には膨大な時間がかかる.

そこで，活用が見込まれる分析アルゴリズムについて高速化技術を研究開発.

-> Grapon(高速な分析アルゴリズム群)

- **クラスタリング，ランキング，推定･分類，グラフ構築, 効率的な並列化など**
- **アルゴリズムの改良で数10倍から数100倍の高速化**
- **1億ノード規模以上のグラフのスケーラブルな処理を目指して取組中.**

＜主な研究成果＞

- 高速Modularityクラスタリング *AAAI'13*
- 高速PageRank *AAAI'13*
- 高速Personalized PageRank *SIGMOD'13*
- 高速Label Propagation *ICML'14*
- 高速グラフ分割（等粒度クラスタリング）*DEIM'15*
- 高速SCAN:Structural Clustering Algorithm for Network *PVLDB(2015)*
- 高速Belief Propagation *IJCAI'15*
- 高速な並列グラフ処理（Rabbit Order） *IPDPS'16*
- 高速L1 sparse graph構築 *PVLDB(2016)*

NTT

グラフマイニング研究に関わる諸活動

価値ある技術の創出と，コラボレーションを通じた検証

データどうしの類似度によるグラフ構築

- グラフを活用した分析例
- グラフ構築手法
- 構築したグラフの概観例

グラフを活用した分析例

Innovative R&D by NTT

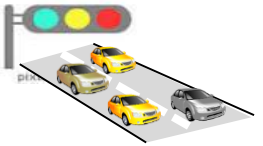
グラフ構造が自明でない場合は，データどうしの類似度からグラフを構築

グラフ構造が自明な場合，その構造を利用（交友関係，道路の接続関係など）

SNSにおけるインフルエンサー発見

交通制御に向けた広域道路網の最適メッシュ分割

予測・信号制御・渋滞回避



グラフ構造が自明でない場合，多次元量から良いグラフを作る必要がある（センサーデータ，画像，文書など）

状態どうしの類類似性を計算し，カテゴリに分類

周りの混雑状況をもとに空いている駐車場を推定

画像どうしの類似性から，画像をカテゴリに分類

路面状態の類似性から正常な路面と劣化した路面とを判別

グラフ構築手法

グラフ構築にはk-NNのほか，全体のバランスを考慮したグラフを構築するものとして，Linear Neighborhood, L1 sparse等が提案されている．

- k-NNグラフ

近傍の上位ｋノードどうしでエッジを張る．

- Linear Neighborhoodグラフ

あるノード$\vec{x_i}$の近傍ノードとの線形和 $\sum W_{ij}\,\vec{x_j}$ を計算．これとの距離が小さくなるように重みW_{ij}を決める．

$$\min\ \varepsilon = \sum_i \left\| \vec{x_i} - \sum_j W_{ij}\,\vec{x_j} \right\|^2$$

$\vec{x_i}$：D x 1　　W_{ij}：重み(スカラ値)　　$\vec{x_j}$：D x 1

凡例
N：　ノード数
D：　次元数

- L1　sparseグラフ

最適化問題を解くことでエッジ有無を自動的に決定．ノイズに強く，データ依存性が少ない．
ただしグラフのノード数よりも（ノードの）次元数が非常に多い場合に有効．

$$\min_{\mathbf{w}_p \in \mathcal{R}^N} \frac{1}{2M} \left\| \vec{\mathbf{x}}_p - \vec{\mathbf{w}}_p \mathbf{X} \right\|_2^2 + \lambda \left\| \vec{\mathbf{w}}_p \right\|_1$$

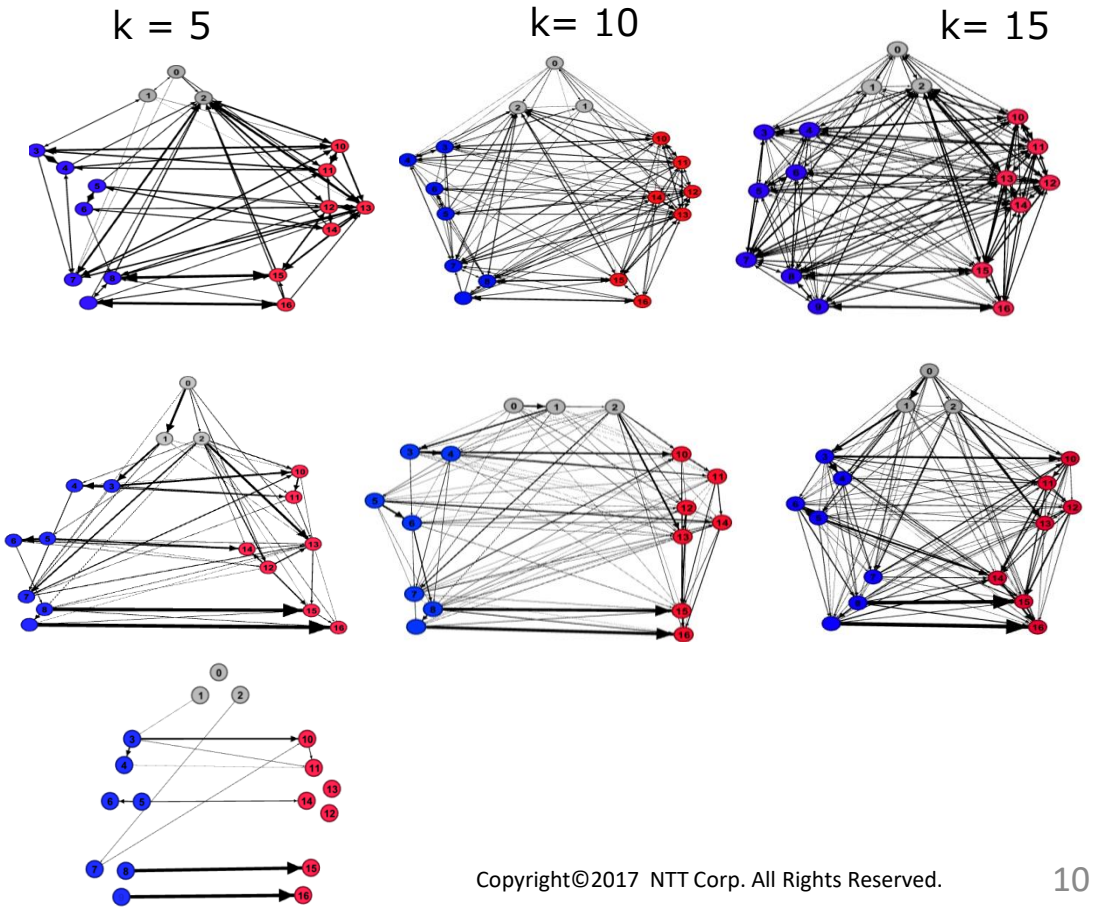
$\vec{\mathbf{x}}_p$：1 x D　　$\vec{\mathbf{w}}_p$：重みベクトル (1 x N)　　$\mathbf{X}$：N x D　　$\vec{\mathbf{w}}_p$：重みベクトルのL1ノルム(ノードpと他ノードとのエッジ重みの絶対値和)

構築したグラフの概観例

Innovative R&D by NTT

Linear NeighboringやL1 sparseは疎なグラフになりやすい傾向.

文書から単語共起による類似度で構築したグラフの例



ラベル伝播による文書分類

- ラベル伝播（Label Propagation）
- ITU勧告文書の分類

ラベル伝播（Label Propagation）　（1/3）

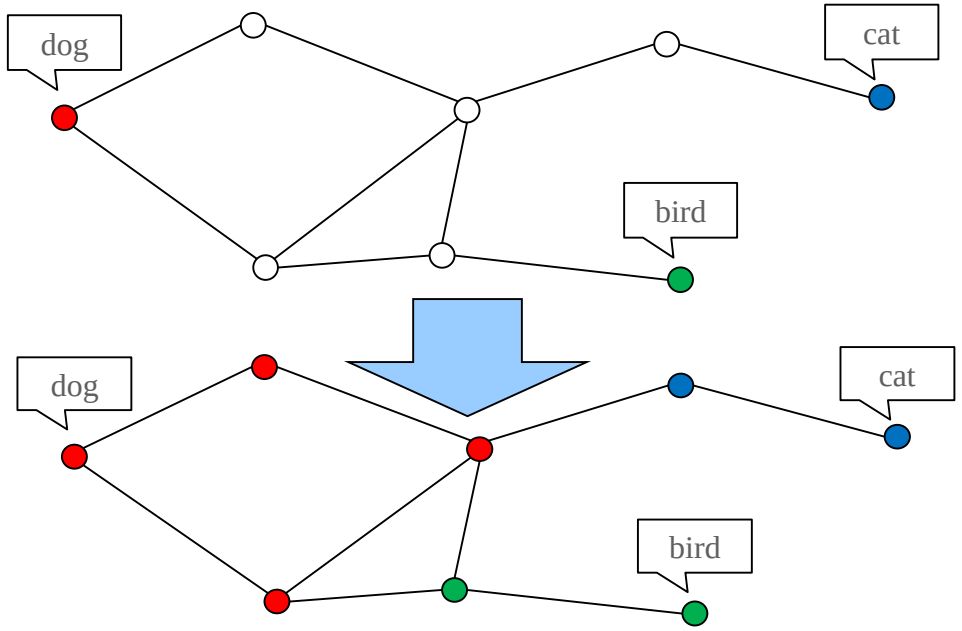
グラフベースの半教師あり学習．ラベル既知のノードから，ラベル未知のノードラベルを推定することで，データのカテゴリを推定・分類する分析に適用可能．

$$F = (I - \alpha S)^{-1} \mathrm{Y}$$

$n \times c$　　$n \times n$　　$n \times c$

n : # of nodes

c : # of category



F : score matrix

I : identity matrix

α : constant parameter

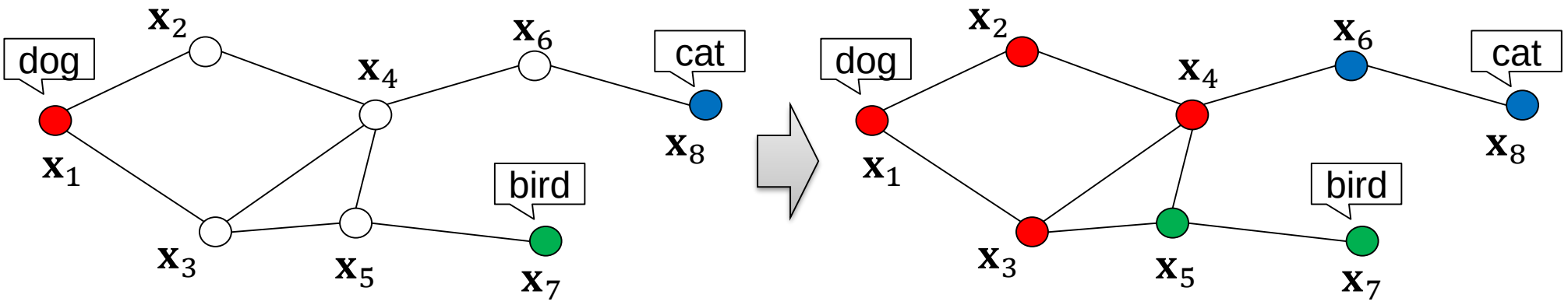
S : normalized adjacency matrix

Y : initial label matrix

ラベル伝播（Label Propagation）　（2/3）

Innovative R&D by NTT

各ラベルのそれぞれについて，スコアを計算していく．



Y^{T}

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
dog	1	0	0	0	0	0	0	0
cat	0	0	0	0	0	0	0	1
bird	0	0	0	0	0	0	1	0

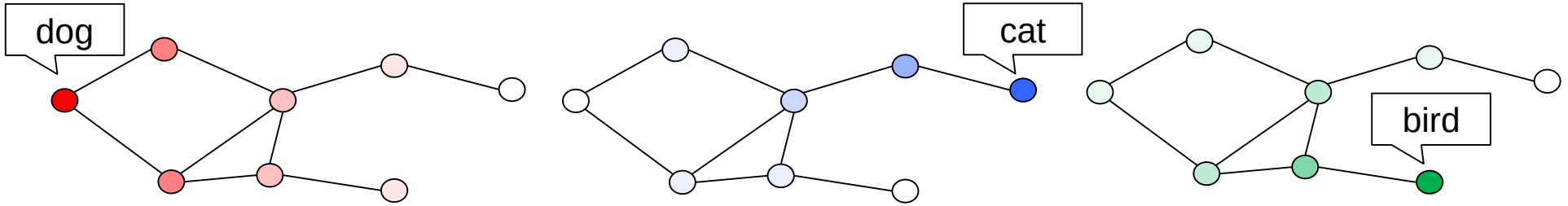
F^{T}

各数値は，候補となる各ラベルの**スコア**

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
dog	1	.9	.8	.6	.1	.1	0	0
cat	0	0	0	.3	0	.9	0	1
bird	0	.1	.2	.1	.9	0	1	0

ラベル伝播（Label Propagation）　（3/3）

- 隣接するノードのラベルどうしは，同じラベルが付与されやすい（smoothness）
- 教師ラベルの近傍は，教師と同じラベルが付与されやすい（fitting）
- αは0~1．0に近いほど教師ラベルの影響を受けやすいく，1に近いほどグラフ構造の影響を受けやすい.



$$\min_F \frac{1}{2}\sum_{i,j}^{n} W_{ij}\left\|\bar{F}_{il}-\bar{F}_{jl}\right\|^2+\left(\frac{1}{\alpha}-1\right)\sum_{i}^{n}\left\|F_{il}-Y_{il}\right\|^2$$

smoothness　　fitting

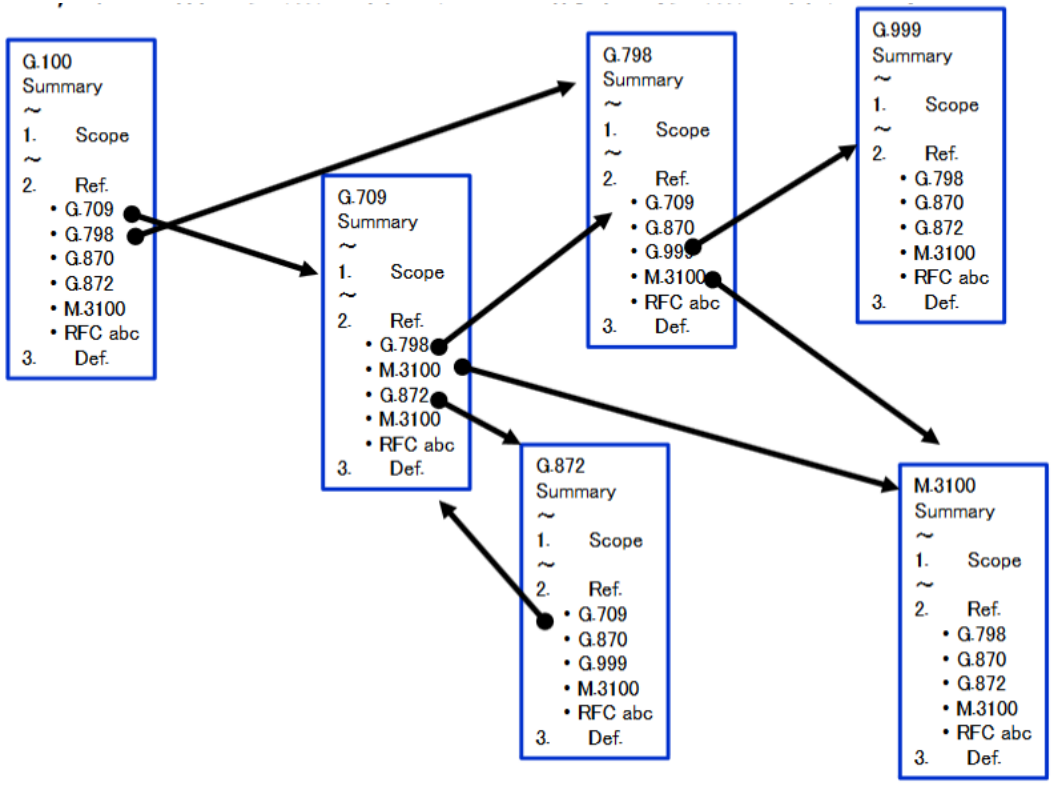
$$\Rightarrow F=(I-\alpha S)^{-1}Y, \qquad S=D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$$

ITU勧告の分類　(1/4)

- **ITU勧告は参照関係が整備されている．クラスタリングにおいて単語共起グラフに比べて参照関係グラフの精度が高いことを確認※**
- **今回，ラベル伝播による分類でも同様かどうかを検証した**

※**モジュラリティ0.62で最も良い精度**．他のグラフ構築手法によるモジュラリティは0.17～0.61（昨年のビッグデータ分析技術ワークショップでご紹介）

参照関係グラフの例



勧告文書の記述例

G.872
Architecture of optical transport networks

Summary

1. Scope
2. Reference
3. Definitions
4. Abbreviations
5. …
6. …

ITU勧告の分類　(2/4)

参照関係のほか，単語共起による類似度でグラフを構築した.

- グラフ構築手法は，k-NN, Linear Neighboring，L1 sparseの3種類

ITU勧告Gシリーズに頻出する単語の例

出現数	word	
17	network	
14	requirement	synchronization
12	equipment packet clock	distribution frequency time
11	timing	
10	layer	
9	Ethernet methods architecture	element IEEE phase physical
8	case protocol	
7	performance signal further minimum study	tolerance details precision PTP support
6	types relevant future specification	application aspects interoperability telecom
5	layers IP media function reference devices	transport slave consistent first functions operation

出現数	word			
4	SDH information different order required currently G.8110 IETF	interfaces limited limits MPLS RFC scenarios particular boundary	conditions environmental generation normal related transfer applicable packet-based	definitions delivery environment G.8260 manner necessary parameters utilize
3	measurements covered jitter G.8261 client end adhered	manufacturers produced quality satisfactory exceeded method accuracy error	addresses noise proper timed base covers form	characteristics configuration design full G.8275 messages profile
2	allocation compliance digital G.709 G.798 individual ODUk optical OTN rate respective structure Additional TDM CES hypothetical metrics	G.803 NEs Option bandwidth connection delay G.8261/Y.1361 holdover interworking PDV period second variation format technology versions NTP	contain level allowed International located national output system detailed framework measurement applies EECOption G.813 G.8271 operates T-GM	G.8265 G.8265/Y.1365 mode carrier general interaction nodes packetbased algorithm best complement exchange G.8275/Y.1369 high-level mapping modes options

ITU勧告の分類　(3/4)

- **分析対象は，ITU-TのGシリーズ(伝送システム及びメディア、ディジタルシステム及びネットワーク)の284件. 参照先も含めて計750件の文書.**
- **これを7カテゴリに分類する. 教師データは全体の10%および30%.**

分析対象のITU勧告文書

カテゴリ	内容	データ数
0	G.100-199 International telephone connection and circuits	25
1	G.200-299 General characteristics common to all analogue carrier-transmission systems	1
2	G.600-699 Transmission media and optical systems characteristics	29
3	G.900-999 Digital sections and digital line system	59
4	G.1000-1999 Multimedia Quality of Service and performance – Generic and user-related aspects	13
5	G.9000-9999 Access networks	21
6	上記以外	136

※カテゴリ分けは，ITU-T標準化活動におけるラポータ（課題担当）経験者のNTTアドバンステクノロジ森田直孝氏によるもの

ITU勧告の分類　(4/4)

- 参照関係のほかは，単語共起(438次元)を用いた
- ラベル伝播は，教師データが少ない場合(10%)に特に他よりも高精度
- 参照関係がもっとも高精度

分類手法	グラフ構築法	正解率				正解率（平均）
		教師データ10%		教師データ30%		
SVM※1	-	57%	69%	84%	82%	73%
K-NN※2	-	43%	55%	66%	67%	58%
ラベル伝播	L1 sparse (α=0.3)	**71%**	**72%**	88%	84%	79%
	L1 sparse (α=0.8)	**77%**	**74%**	89%	86%	82%
	Linear Neighborhood (α=0.3)	**63%**	**73%**	82%	83%	75%
	Linear Neighborhood (α=0.8)	**65%**	**77%**	81%	82%	70%
	K-NN (α=0.3)	**61%**	**65%**	77%	77%	70%
	K-NN (α=0.8)	**67%**	**64%**	71%	70%	68%
	参照関係 (α=0.3)	**80%**	**81%**	91%	91%	**85%**
	参照関係 (α=0.8)	**80%**	**81%**	91%	90%	**85%**

※1 SVC with linear kernel
※2 K=5. ユークリッド距離

大規模化対応

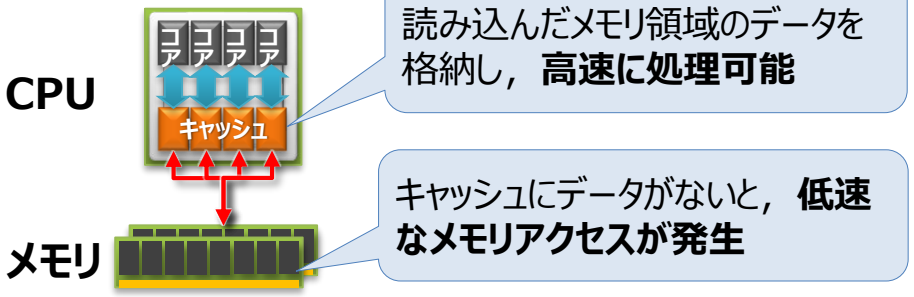
- グラフのReordering
- Reorderingによる効率的な並列処理
- 並列処理によるラベル伝播の高速化例

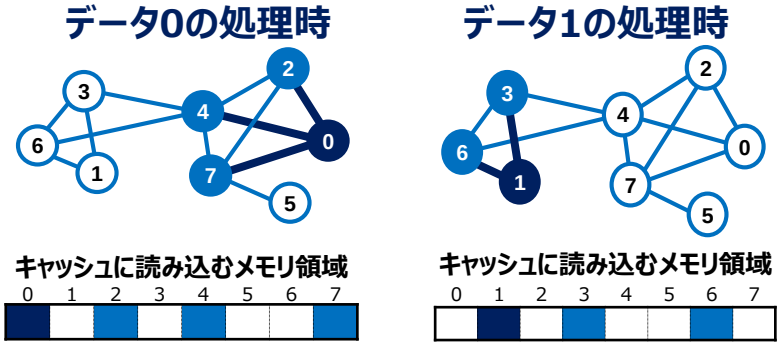
グラフのReordering

データのID番号の並びを最適化(reorder)してキャッシュヒット率を多くする.

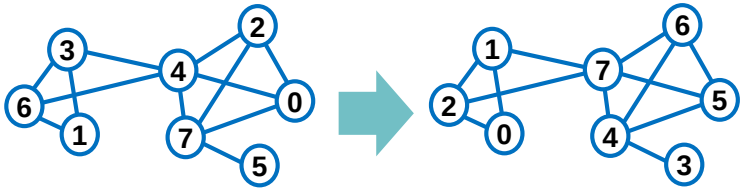
グラフマイニングにおけるメモリアクセス

隣接するデータを順にメモリから読み込んでいく

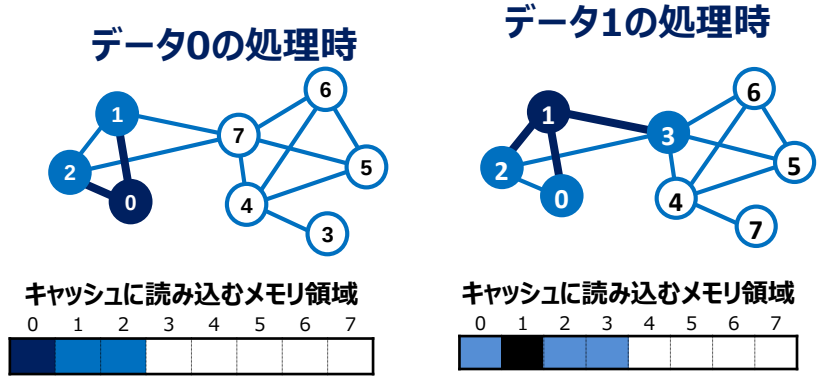




本技術: データの並びを最適化



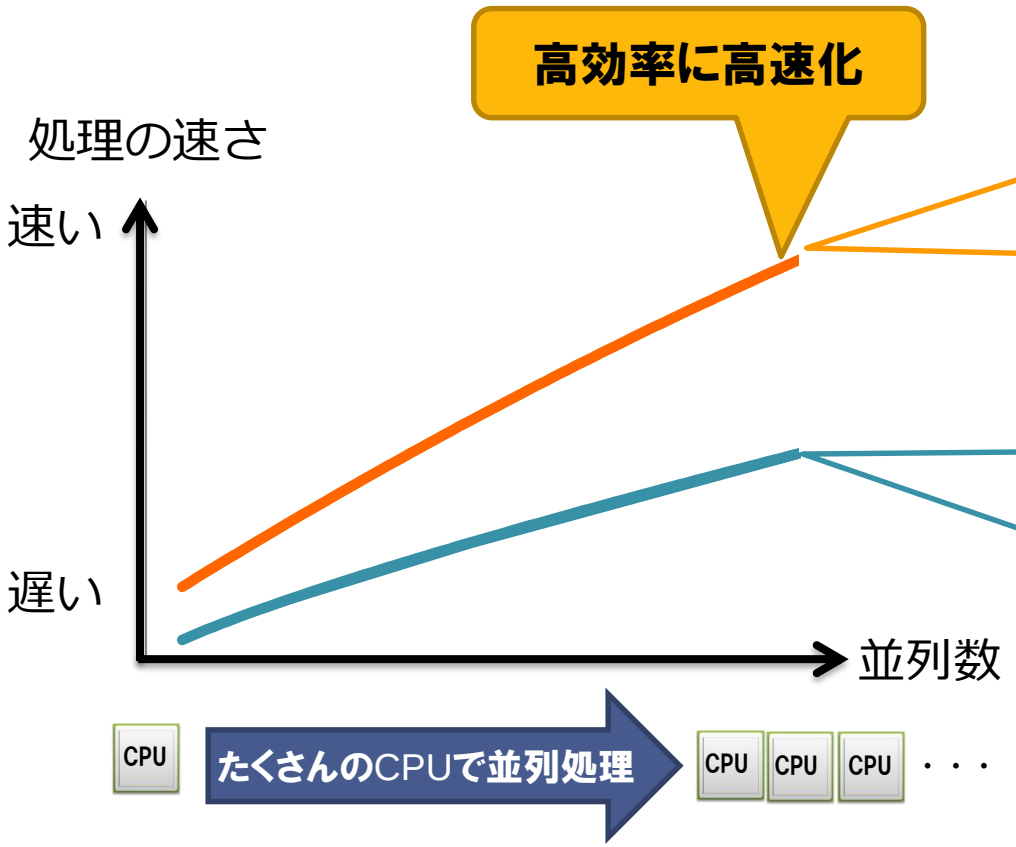
データのID番号を振りなおし，隣接するデータどうしをメモリ上の近傍に配置



キャッシュを最大限活用
不要なメモリアクセスとCPU間通信を低減

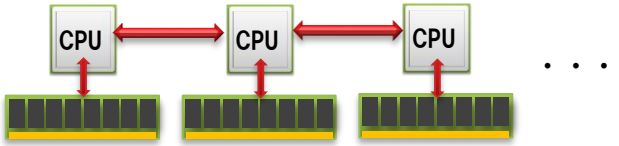
Reorderingによる効率的な並列処理

グラフ処理の並列化により，要件に合わせた高速化を可能とする．
メモリバンド幅のボトルネックを回避する高いスケール性能を目指す．



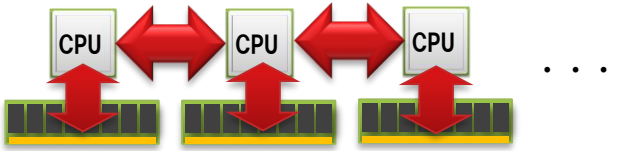
Reorderによる効率的な並列処理

メモリアクセスとCPU間通信を抑えられ，並列数にあわせたスケールが可能



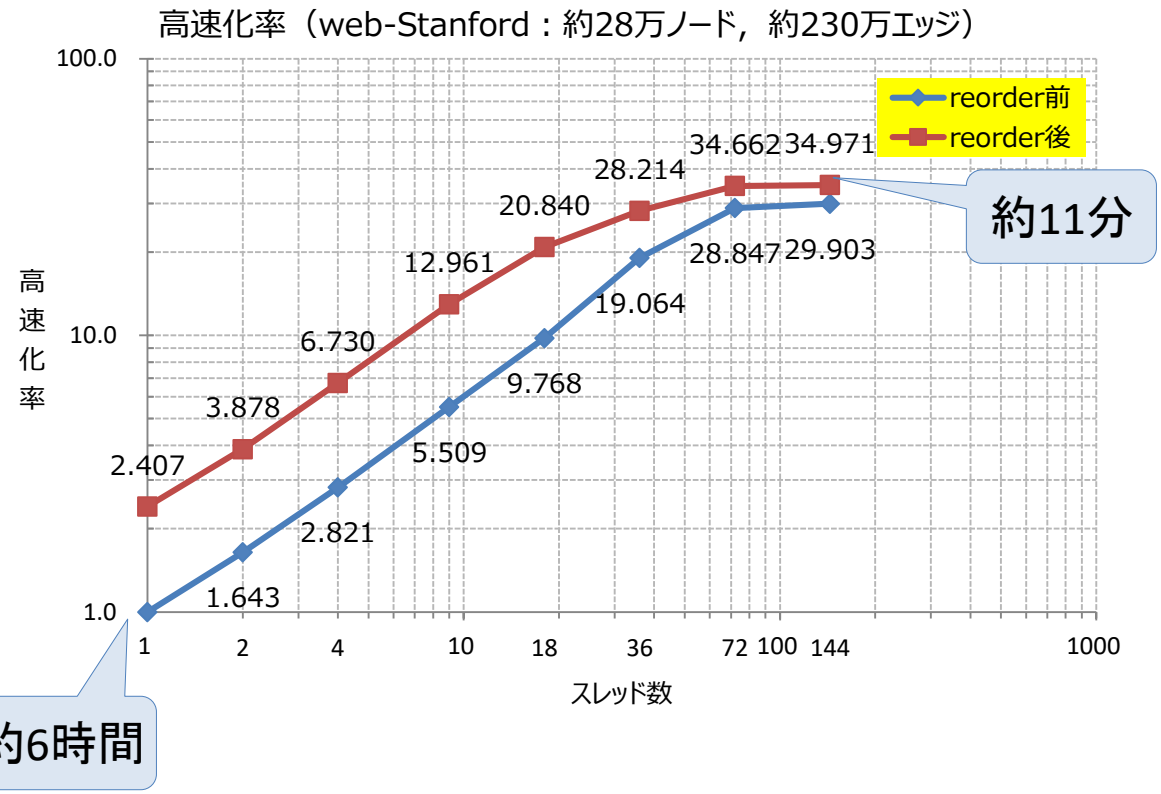
グラフ処理の並列化（reorder前）

ただし，このままではメモリアクセスとCPU間通信が多く発生し，なかなかスケールしない可能性



並列処理によるラベル伝播の高速化例

- Web-Stanfordデータについて，72コア環境で35倍高速化
- 並列化にはOpenMPを利用



測定環境：Xeon E7-8890v3 (物理18コア) x 4ソケット / メモリ2TB

まとめと今後の予定

まとめ

- グラフマイニングの適用例として，ITU勧告文書群からグラフを構築し，ラベル伝播を行うことで，少ない教師データで高精度に分類できることを示した.

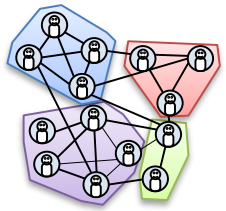
- グラフマイニングの大規模化対応として，効率的な並列処理を行うことでラベル伝播を高効率にスケール可能であることを示した.

今後の予定

多様な属性を持つグラフデータを分析可能とし，さまざまな現実の課題解決につなげる.

ソーシャルグラフ

属性を無視した同質なノードとエッジに縮約
エッジは直感的なものを利用



ノード：人（のみ）
エッジ：交友関係

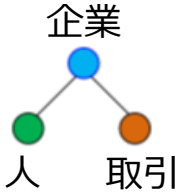
さまざまなセンサからのデータ分析・処理

たとえば，データをグラフ構造として捉えることで効率的な処理ができる分野の開拓を目指す

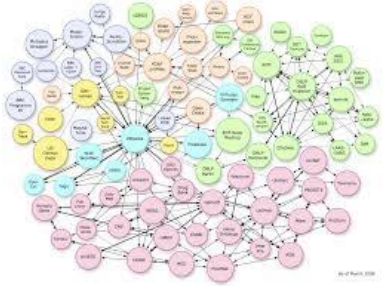
確率伝播による領域分割

プロパティグラフからの検索・マイニング

たとえば，属性を考慮することで，多様な類似構造の発見・検索等のタスクを扱うことを目指す

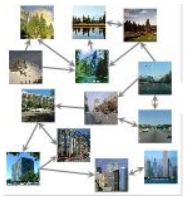


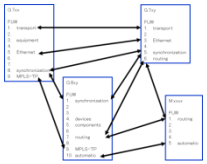
Linked Open Data



画像どうしの類似度など多様なグラフ構造をマイニング

たとえば、グラフ構造を与えることで，教師データが少ない場合でも高精度にクラスタリング・分類するタスクを扱うことを目指す





Thank you

Efficient Mining Algorithms for Large-scale Graphs
NTT Technical Review, Dec 2013, Vol. 11, No. 12

Advanced Processing and Analytics for Large-scale Graphs
NTT Technical Review, Feb 2016, Vol. 14, No. 2

Grapon

NTT