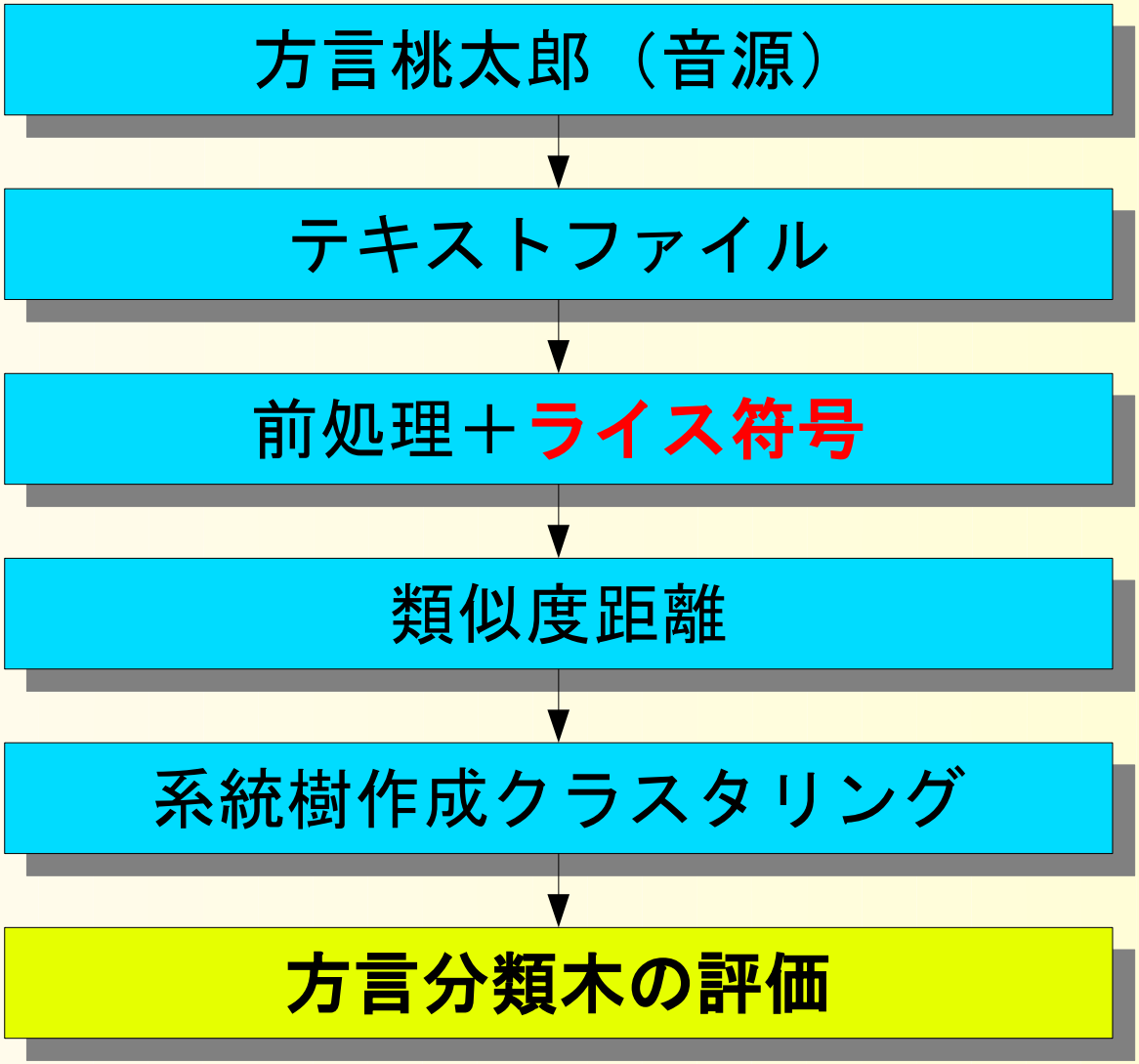
圧縮類似度を用いた方言の自動分類
～ライス符号を用いた前処理～
～連結クラスタリング法～
～余弦類似度を用いた方言分類木の評価～

日本大学文理学部
情報システム解析学科

谷　研究室　　本荘　智則
佐藤　雄太
益田　真太郎

本研究の手順



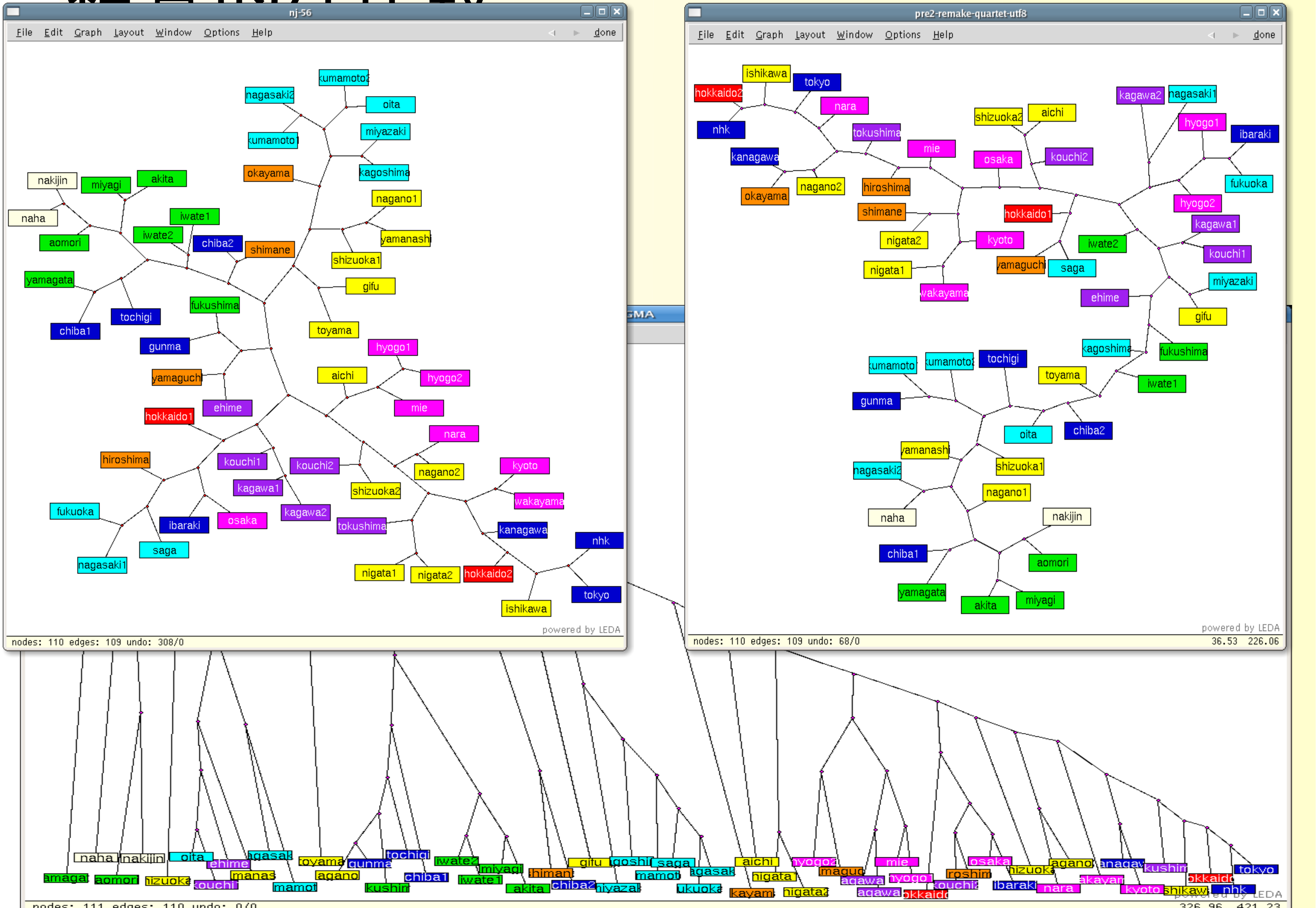
目次

目次

背景

NJ法 UPGMA法 Quartet Methodの
3種類でグラフを作成
3種類のどれがより良い分類をしているだろうか？？？

視覚的な比較

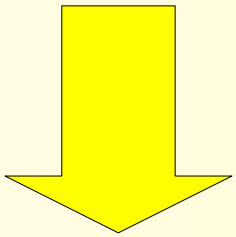

nj-56
File Edit Graph Layout Window Options Help done
powered by LEDA
nodes: 110 edges: 109 undo: 308/0
pre2-remake-quartet-utf8
File Edit Graph Layout Window Options Help done
powered by LEDA
nodes: 110 edges: 109 undo: 68/0
36.53 226.06
nodes: 111 edges: 110 undo: 0/0
326.96 421.23

背景

3 種類のどれが木として
より類似度距離を反映しているか
グラフを見ても客観的には分からない

背景

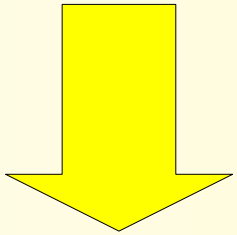
3種類のどれが木として
より類似度距離を反映しているか
グラフを見ても客観的には分からない



3種類のグラフでどれが木として類似度距離をより反映しているか？

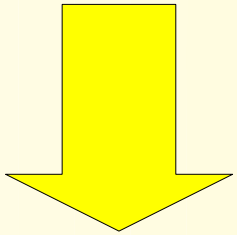
背景

3種類のグラフでどれが木として類似度距離をより反映しているか？



背景

3種類のグラフでどれが木として類似度距離をより反映しているか？



反映しているか評価する値を定義し
客観的な木の評価を行う

背景

3種類のグラフでどれが類似度距離をより反映しているか？？

2007年度谷研究室在籍の堀中氏の先行研究

反映しているか評価する値を定義し

客観的な木の評価を行う

目次

先行研究．木の距離定義

先行研究．木の距離定義

グラフを見た場合の距離

sd(u, v)：類似度距離とする
td(u, v) を類似度の距離に対応させた距離を
ntd(u, v)　とする

td(u, v)：グラフ間の
ノードの辺の数を距離と
したもの

先行研究．木の距離定義

グラフを見た場合の距離

sd(u, v)：類似度距離とする
td(u, v) を類似度の距離に対応させた距離を
ntd(u, v)　とする

td(u, v)：グラフ間のノードの辺の数を距離としたもの

なぜ ntd(u, v) を定義する？

先行研究．木の距離定義

グラフを見た場合の距離

sd(u, v)：類似度距離とする
td(u, v) を類似度の距離に対応させた距離を
ntd(u, v)　とする

td(u, v)：グラフ間の
ノードの辺の数を距離と
したもの

なぜ ntd(u, v) を定義する？

td(u, v)：グラフで定まる　sd(u, v)：距離表で定まる

先行研究．木の距離定義

グラフを見た場合の距離

sd(u, v)：類似度距離とする
td(u, v) を類似度の距離に対応させた距離を
ntd(u, v)　とする

td(u, v)：グラフ間の
ノードの辺の数を距離と
したもの

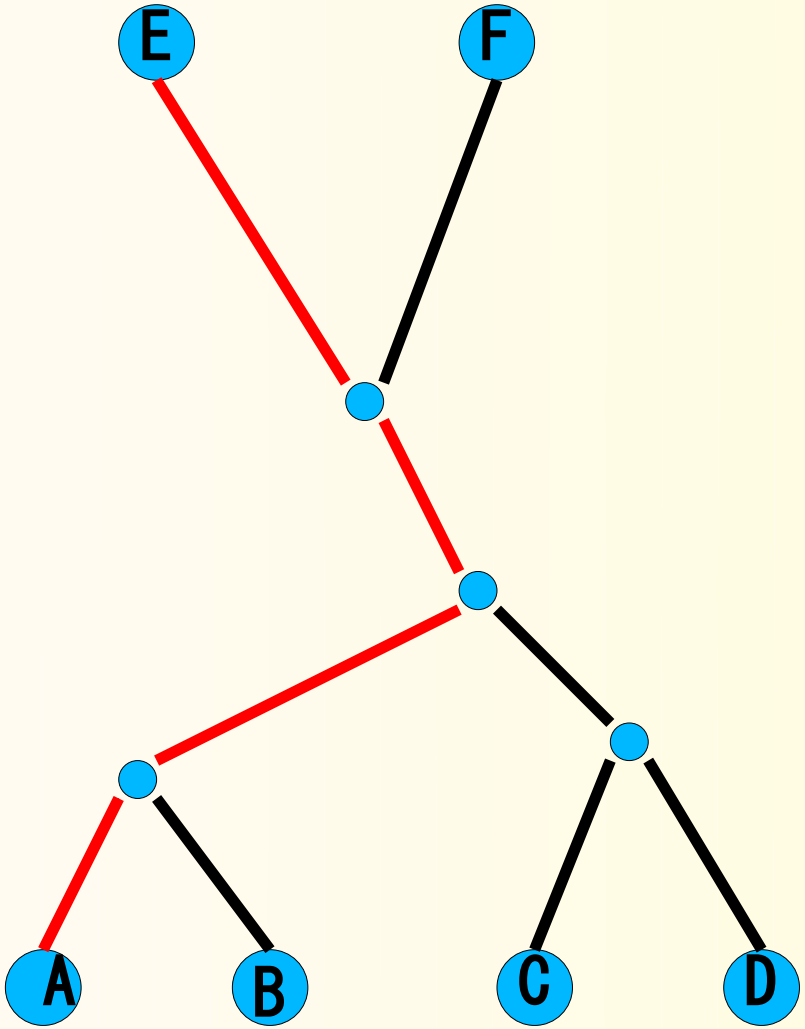
なぜ ntd(u, v) を定義する？

td(u, v)：グラフで定まる　sd(u, v)：距離表で定まる

木における距離 td(u, v) を類似度距離 sd(u, v) に対応させることで
木が類似度距離にどれ位反映しているか検証できるから

先行研究．木の距離定義

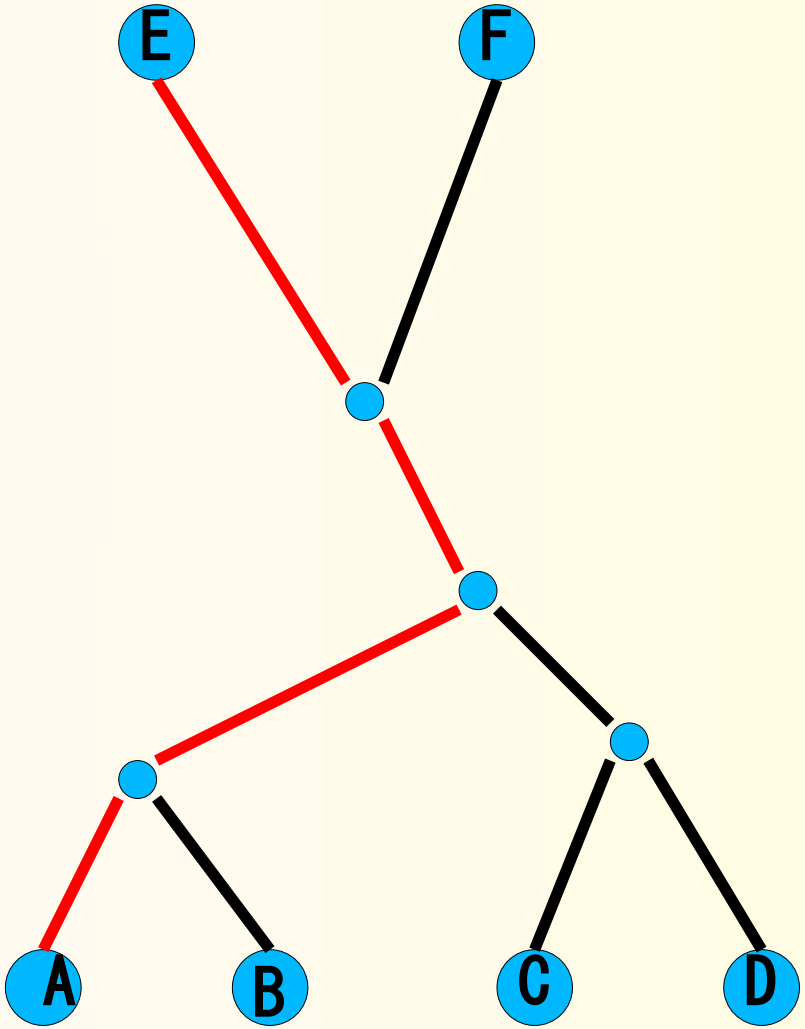
グラフを見た場合の距離



td(u, v)：グラフ間のノードの辺の数を距離としたもの

先行研究．木の距離定義

グラフを見た場合の距離



td(u, v)：グラフ間のノードの辺の数を距離としたもの

td(A, E)=4

先行研究．木の距離定義

グラフを見た場合の距離

td(u, v) を類似度の距離に対応させた距離を ntd(u, v) とする

先行研究．木の距離定義

グラフを見た場合の距離

td(u, v) を類似度の距離に対応させた距離を
ntd(u, v)　とする

sd(u, v)：類似度距離

先行研究．木の距離定義

グラフを見た場合の距離

td(u,v) を類似度の距離に対応させた距離を
ntd(u,v)　とする

sd(u,v)：類似度距離

max(sd)：sd(u,v) の最大値
min(sd)：sd(u,v) の最小値
max(td)：td(u,v) の最大値
min(td)：td(u,v) の最大値

先行研究．木の距離定義

グラフを見た場合の距離

td(u,v) を類似度の距離に対応させた距離を
ntd(u,v)　とする

sd(u,v)：類似度距離

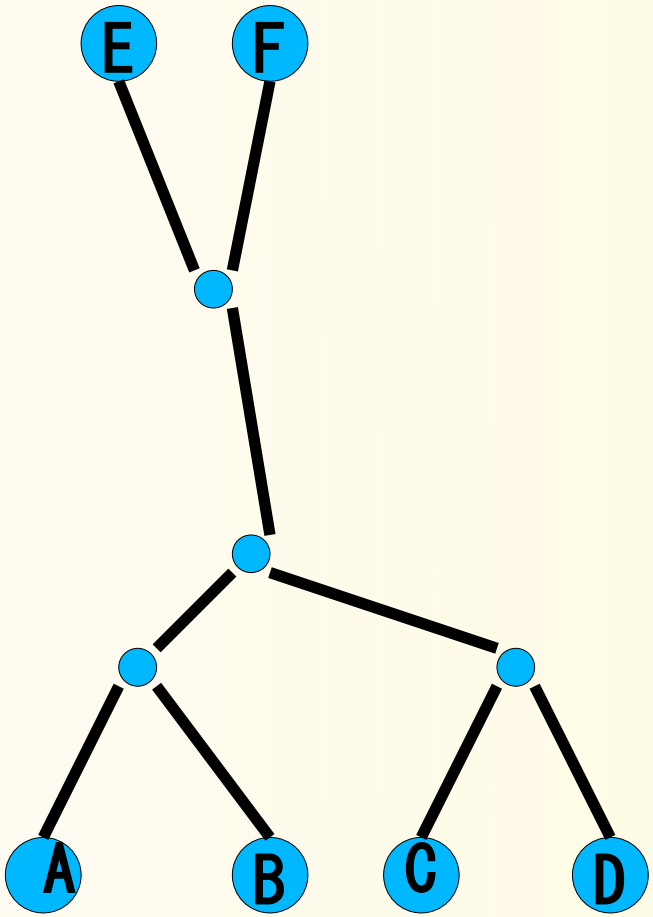
max(sd)：sd(u,v) の最大値
min(sd)：sd(u,v) の最小値
max(td)：td(u,v) の最大値
min(td)：td(u,v) の最大値

$$ntd(u,v) = min(sd) + \frac{max(sd)-min(sd)}{max(td)-min(td)} (td(u,v)-min(td))$$

先行研究．木の距離定義

グラフを見た場合の距離

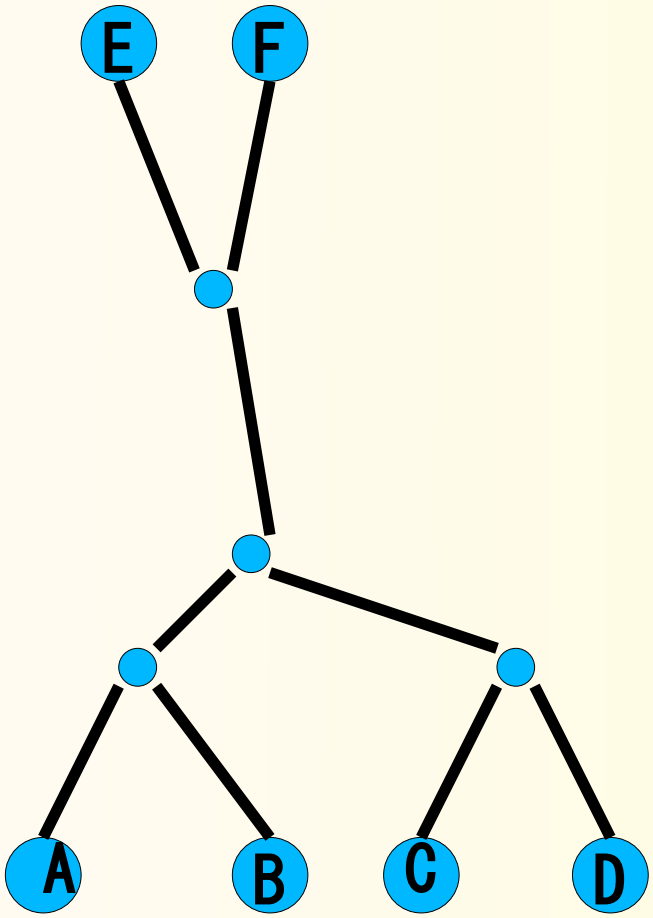
td(u, v) を類似度距離に対応させた距離を
ntd(u, v) とする



先行研究．木の距離定義

グラフを見た場合の距離

td(u,v) を類似度距離に対応させた距離を ntd(u,v) とする



max(sd):17

min(sd):5

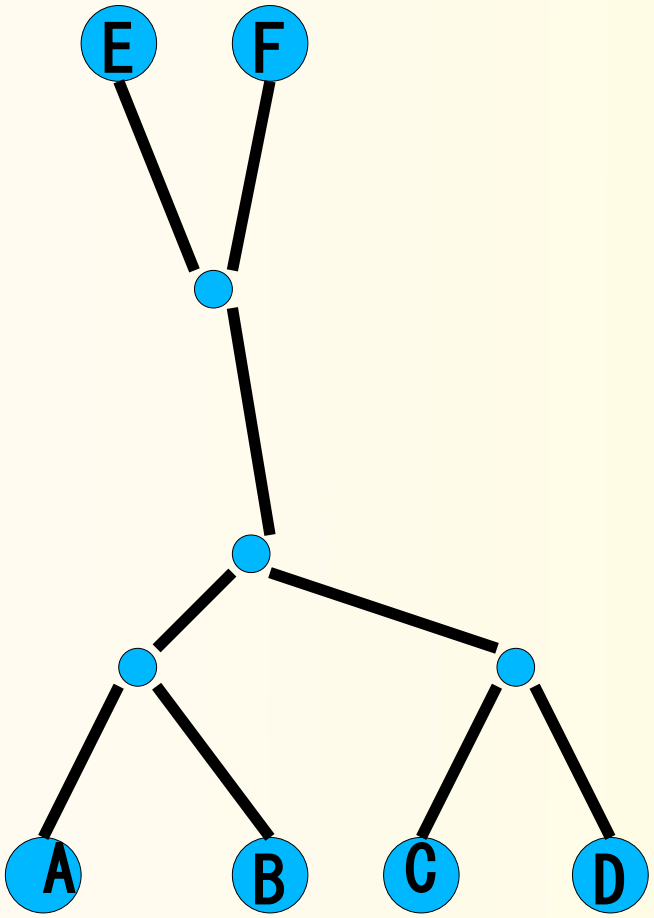
max(td):4

min(td):2

先行研究．木の距離定義

グラフを見た場合の距離

td(u,v) を類似度距離に対応させた距離を ntd(u,v) とする



max(sd):17

min(sd):5

max(td):4

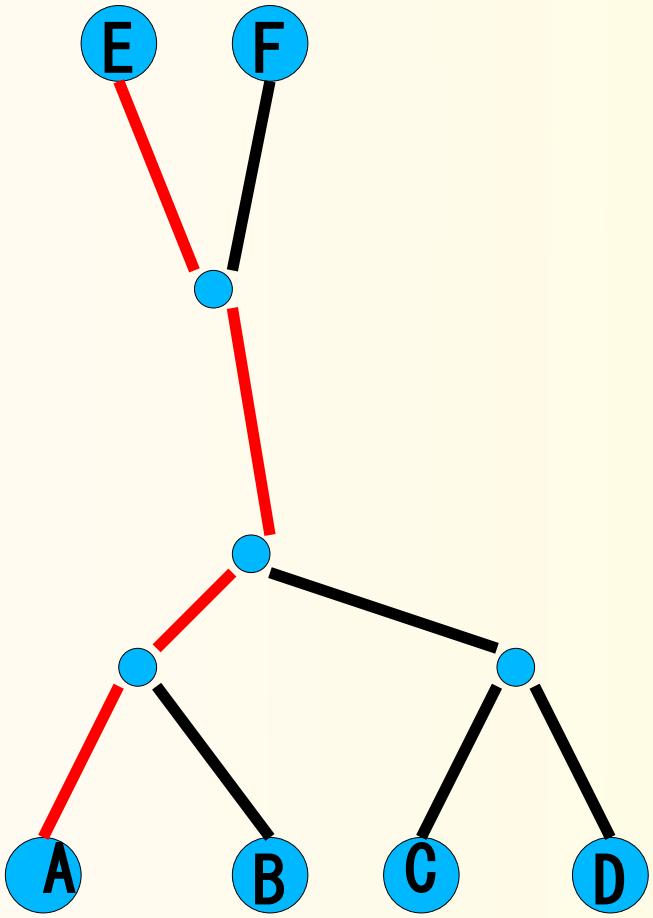
min(td):2

$$\frac{\max(sd)-\min(sd)}{\max(td)-\min(td)} = \frac{12}{2} = 6$$

先行研究．木の距離定義

グラフを見た場合の距離

td(u,v) を類似度距離に対応させた距離を ntd(u,v) とする



max(sd):17

min(sd):5

max(td):4

min(td):2

ntd(A,E):5+6(4-2)=5+12=17

先行研究．木の評価の定義値

$$S(T) = \frac{(M - C_T)}{(M - m)}$$

S(T) 値

$$TV_1 = \frac{2}{N(N-1)} \sum |sd(u,v) - ntd(u,v)|$$

平均値

$$TV_2 = \sqrt{\sum (sd(u,v) - ntd(u,v))^2}$$

2 乗和の平方根

sd(u, v)：類似度距離

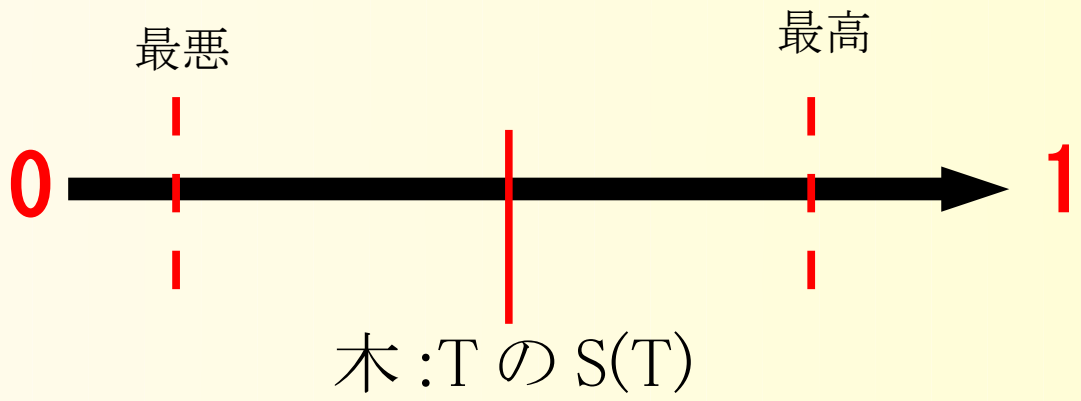
先行研究．{S(T) 値}

$$S(T) = \frac{(M - C_T)}{(M - m)}$$

M :maximum cost

m :minimum cost

C_T :total cost



先行研究．木の評価の定義値

$$S(T) = \frac{(M - C_T)}{(M - m)}$$

S(T) 値

$$TV_1 = \frac{2}{N(N-1)} \sum |sd(u,v) - ntd(u,v)|$$

平均値

$$TV_2 = \sqrt{\sum (sd(u,v) - ntd(u,v))^2}$$

2 乗和の平方根

sd(u, v)：類似度距離

TV は 値が 0 に，S(T) は 1 に近い程 sd(u, v) をより反映したグラフ

先行研究．{S(T) 値}

$$S(T) = \frac{(M - C_T)}{(M - m)}$$

		value
Preprocess2	Quartet Method	0.850286
	NJ	0.455565
	UPGMA	0.455565

先行研究.（差の平均）

$$TV_1 = \frac{2}{N(N-1)} \sum |sd(u,v) - ntd(u,v)|$$

先行研究.（差の平均）

$$TV_1 = \frac{2}{N(N-1)} \sum |sd(u,v) - ntd(u,v)|$$

		value
Preprocess2	NJ	0.130395
	UPGMA	0.162626
	Quartet Method	0.177194

先行研究 . (2 乗和の平方根)

$$TV_2 = \sqrt{\sum (sd(u,v) - ntd(u,v))^2}$$

先行研究．(2 乗和の平方根)

$$TV_2 = \sqrt{\sum (sd(u,v) - ntd(u,v))^2}$$

		value
Preprocess2	NJ	6.352560
	UPGMA	7.745910
	Quartet Method	8.268810

先行研究．結果

$$S(T) = \frac{(M - C_T)}{(M - m)}$$

		value
Preprocess2	Quartet Method	0.850286
	NJ	0.455565
	UPGMA	0.455565

$$TV_1 = \frac{2}{N(N-1)} \sum |sd(u,v) - ntd(u,v)|$$

		value
Preprocess2	NJ	0.130395
	UPGMA	0.162626
	Quartet Method	0.177194

先行研究．結果

S(T) =

Quartet-Methodは sd(u,v) をあまり反映しない

		value
Preprocess2	Quartet Method	0.850286
	NJ	0.455565
	UPGMA	0.455565

$$TV_1 = \frac{2}{N(N-1)} \sum |sd(u,v) - ntd(u,v)|$$

		value
Preprocess2	NJ	0.130395
	UPGMA	0.162626
	Quartet Method	0.177194

先行研究．結果

S(T) =

Quartet-Method は sd(u, v) をあまり反映しない

		value
Preprocess2	Quartet Method	0.850286
	NJ	0.455565
	UPGMA	0.455565

TV_1 =

TV は NJ 法がよかった

		value
Preprocess2	NJ	0.130395
	UPGMA	0.162626
	Quartet Method	0.177194

目次

研究動機

情報が失われず信頼できる評価方法はないか？？

研究動機

情報が失われず信頼できる評価方法はないか？？

ベクトルで類似度判定を行う余弦類似度

目次

研究概要

本研究

・余弦類似度を用いた分類木の新たな評価法

以上から木の評価の精度をあげる

研究概要

本研究

・余弦類似度を用いた分類木の新たな評価法

以上から木の評価の精度をあげる

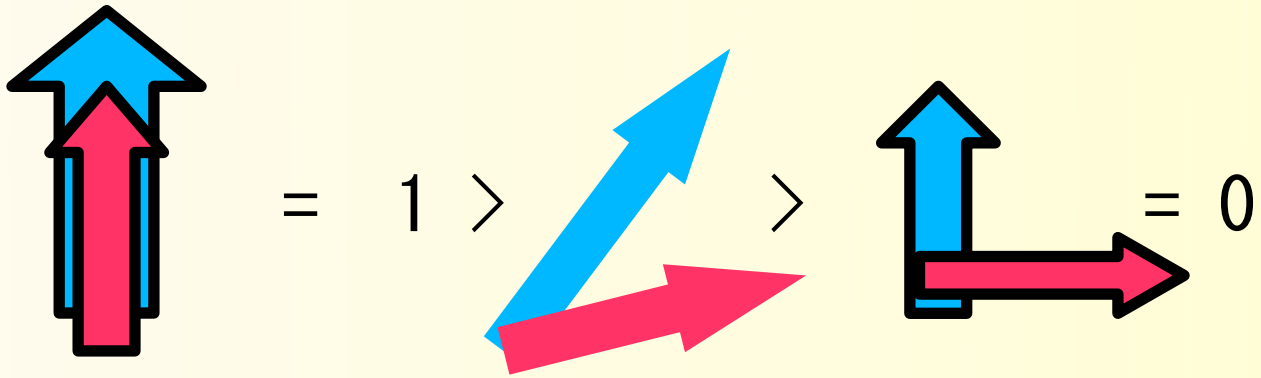
・従来の前処理のデータと符号化されたデータの比較

目次

余弦類似度

ベクトルｐとベクトルｑの余弦を計算して
２つのベクトルの類似度で評価する

$$\mathrm{COS}(\mathbf{p},\mathbf{q}) = \frac{\langle \mathbf{p},\ \mathbf{q} \rangle}{|\mathbf{p}| \cdot |\mathbf{q}|}$$



余弦類似度

ベクトルｐとベクトルｑの余弦を計算して
２つのベクトルの類似度で評価する

$$COS(\mathbf{p},\mathbf{q}) = \frac{\langle \mathbf{p},\ \mathbf{q} \rangle}{|\mathbf{p}| \cdot |\mathbf{q}|}$$

入力： sd 距離表 , ntd 距離表

出力：　　０から１の値

目次

木の評価．定義値

$$S(T) = \frac{(M - C_T)}{(M - m)}$$

S(T) 値

$$TV_1 = \frac{2}{N(N-1)} \Sigma \ |sd(u,v) - ntd(u,v)|$$

平均値

$$TV_2 = \sqrt{\sum (sd(u,v) - ntd(u,v))^2}$$

2 乗和の平方根

木の評価．定義値

$$S(T) = \frac{(M - C_T)}{(M - m)}$$

S(T) 値

$$TV_1 = \frac{2}{N(N-1)} \Sigma \ |sd(u,v) - ntd(u,v)|$$

平均値

$$TV_2 = \sqrt{\sum (sd(u,v) - ntd(u,v))^2}$$

2 乗和の平方根

$$CV : \cos\theta = \frac{\langle sd(u,v) \cdot ntd(u,v) \rangle}{||sd(u,v)|| \cdot ||ntd(u,v)||}$$

余弦類似度

木の評価．定義値

$$S(T) = \frac{(M - C_T)}{(M - m)}$$

S(T) 値

$$TV_1 = \frac{2}{N(N-1)} \Sigma \ |sd(u,v) - ntd(u,v)|$$

平均値

$$TV_2 = \sqrt{\sum (sd(u,v) - ntd(u,v))^2}$$

2 乗和の平方根

$$CV : \cos\theta = \frac{\langle sd(u,v) \cdot ntd(u,v) \rangle}{||sd(u,v)|| \cdot ||ntd(u,v)||}$$

余弦類似度

TV は 値が 0 に，S(T)，CV は 1 に近い程 sd をより反映したグラフ

目次

3.　符号化無し {S(T) 値 }

$$S(T) = \frac{(M - C_T)}{(M - m)}$$

	preprocess1	preprocess2
NJ 法	0.487411	0.482618
UPGMA	0.487411	0.482618
Quartet-Method	0.347207	0.632892

3. 符号化無し（差の平均）

$$TV_1 = \frac{2}{N(N-1)} \sum |sd(u,v) - ntd(u,v)|$$

	preprocess1	preprocess2
NJ 法	0.106931	0.121357
UPGMA	0.153092	0.151837
Quartet-Method	0.294744	0.275692

3.　符号化無し（2乗和の平方根）

$$TV_2 = \sqrt{\sum (sd(u,v) - ntd(u,v))^2}$$

	preprocess1	preprocess2
NJ 法	4.98775	5.64342
UPGMA	6.42743	6.36746
Quartet-Method	13.8222	13.2202

3.　符号化無し（余弦類似度）

$$CV : \cos\theta = \frac{\langle sd(u,v) \cdot ntd(u,v) \rangle}{||sd(u,v)|| \cdot ||ntd(u,v)||}$$

	preprocess1	preprocess2
NJ 法	0.472792	0.494139
UPGMA	0.476677	0.49831
Quartet-Method	0.0254528	0.0270263

3. 符号化 {S(T) 値}

$$S(T) = \frac{(M - C_T)}{(M - m)}$$

	preprocess1	preprocess1	preprocess2	preprocess2
	b=8	b=16	b=8	b=16
NJ 法	0.503624	0.508153	0.49763	0.474975
UPGMA	0.503624	0.508153	0.49763	0.474975
Quartet-Method	0.828499	0.811279	0.819522	0.798682

3. 符号化（差の平均）

$$TV_1 = \frac{2}{N(N-1)} \sum |sd(u,v) - ntd(u,v)|$$

	preprocess1	preprocess1	preprocess2	preprocess2
	b=8	b=16	b=8	b=16
NJ 法	0.643953	0.051637	0.0610053	0.831602
UPGMA	0.140783	0.0992494	0.199053	0.156328
Quartet-Method	0.803195	0.902535	0.855343	0.888123

3.　符号化（2乗和の平方根）

$$TV_2 = \sqrt{\sum (sd(u,v) - ntd(u,v))^2}$$

	preprocess1	preprocess1	preprocess2	preprocess2
	b=8	b=16	b=8	b=16
NJ 法	3.12369	2.8588	4.0197	4.03721
UPGMA	5.69323	4.18645	8.14697	6.56131
Quartet-Method	32.6971	36.8786	34.7177	36.2311

3. 符号化（余弦類似度）

$$CV : \cos\theta = \frac{\langle sd(u,v) \cdot ntd(u,v) \rangle}{||sd(u,v)|| \cdot ||ntd(u,v)||}$$

	preprocess1	preprocess1	preprocess2	preprocess2
	b=8	b=16	b=8	b=16
NJ 法	0.600193	0.849169	0.597294	0.626059
UPGMA	0.602148	0.850227	0.628886	0.762099
Quartet-Method	0.032723	0.0410279	0.0323438	0.0385313

3. 結果一覧（符号化無し）

	preprocess1	preprocess1	preprocess1	preprocess1
	S(T)	TV1	TV2	CV
NJ 法	0.427187	0.15427	1.8345	0.351619
UPGMA	0.427187	0.245212	16.8321	0.0168566
Quartet-Method	0.41834	0.256335	12.4018	0.0199777

	preprocess2	preprocess2	preprocess2	preprocess2
	S(T)	TV1	TV2	CV
NJ 法	0.351794	0.15508	1.85215	0.37094
UPGMA	0.351794	0.21914	15.5783	0.0178468
Quartet-Method	0.504066	0.248758	12.389	0.0181531

3. 結果一覧（符号化－前処理 1）

rice/b=8	preprocess1	preprocess1	preprocess1	preprocess1
	S(T)	TV1	TV2	CV
NJ 法	0.576013	0.0643147	0.727892	0.50471
UPGMA	0.576013	0.730638	29.7128	0.0245355
Quartet-Method	0.472181	0.294063	13.8375	0.0273412

rice/b=16	preprocess1	preprocess1	preprocess1	preprocess1
	S(T)	TV1	TV2	CV
NJ 法	0.465594	0.0573145	0.693105	0.822158
UPGMA	0.465594	0.282483	20.4416	0.0392706
Quartet-Method	0.579318	0.372329	17.5248	0.0446572

3. 結果一覧（符号化－前処理2）

rice/b=8	preprocess2	preprocess2	preprocess2	preprocess2
	S(T)	TV1	TV2	CV
NJ 法	0.49459	0.0933362	1.12188	0.725118
UPGMA	0.49459	0.271887	19.6051	0.0345233
Quartet-Method	0.561546	0.360401	16.9093	0.0385271

rice/b=16	preprocess2	preprocess2	preprocess2	preprocess2
	S(T)	TV1	TV2	CV
NJ 法	0.439778	0.0712431	0.808413	0.847463
UPGMA	0.439778	0.281534	20.4286	0.0402256
Quartet-Method	0.61914	0.390121	18.4503	0.0414795

目次

4.　考察

1. 余弦類似度でも2007年度研究と同じで
   NJ法の評価値が一番よかった
   ・過去の研究から３つののクラスタリングの中では
   　NJ法がより類似度距離を反映すると考えられるため
2. 前処理２の評価値が前処理１よりよい
   ・50音順より使用頻度順のデータの方が圧縮後の
   　ファイルサイズが小さくなったと考えられるため
3. 符号化すると評価値がよくなった
   ・符号化することにより圧縮前のファイルサイズが小さく
   　なったと考えられるため
4. ライス符号のb=16のとき評価が最もよくなった
   ・パラメータを大きくしたことにより
   　更に圧縮前のファイルサイズが小さくなったと考えられるため

4.　今後の課題

Nj法　UPGMA法 Quartet Methodの
系統樹作成クラスタリング以外でも
評価値を出す
（例：連結クラスタリング）

- 御清聴ありがとうございます

1. はじめに－先行研究

過去の入力

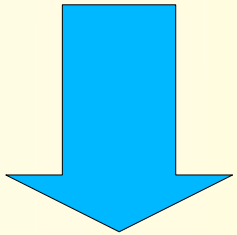
2007 年　音声ファイルを人が聞いてそれを手作業でテキスト化したもの
・とてもよい結果が出たが、以下の２点の問題があった
－テキスト化に人間が介入している
－方言で重要なはずの音声情報を全て捨てている

2008 年　テキスト化に「ドラゴンスピーチ」ソフトを使い自動的にテキスト化したもの
・ 2007 年研究より良くない結果となった

2009 年　テキストのみのデータに加え音声情報（ピッチを付加したもの）
・ 2007 年研究より良くない結果となった

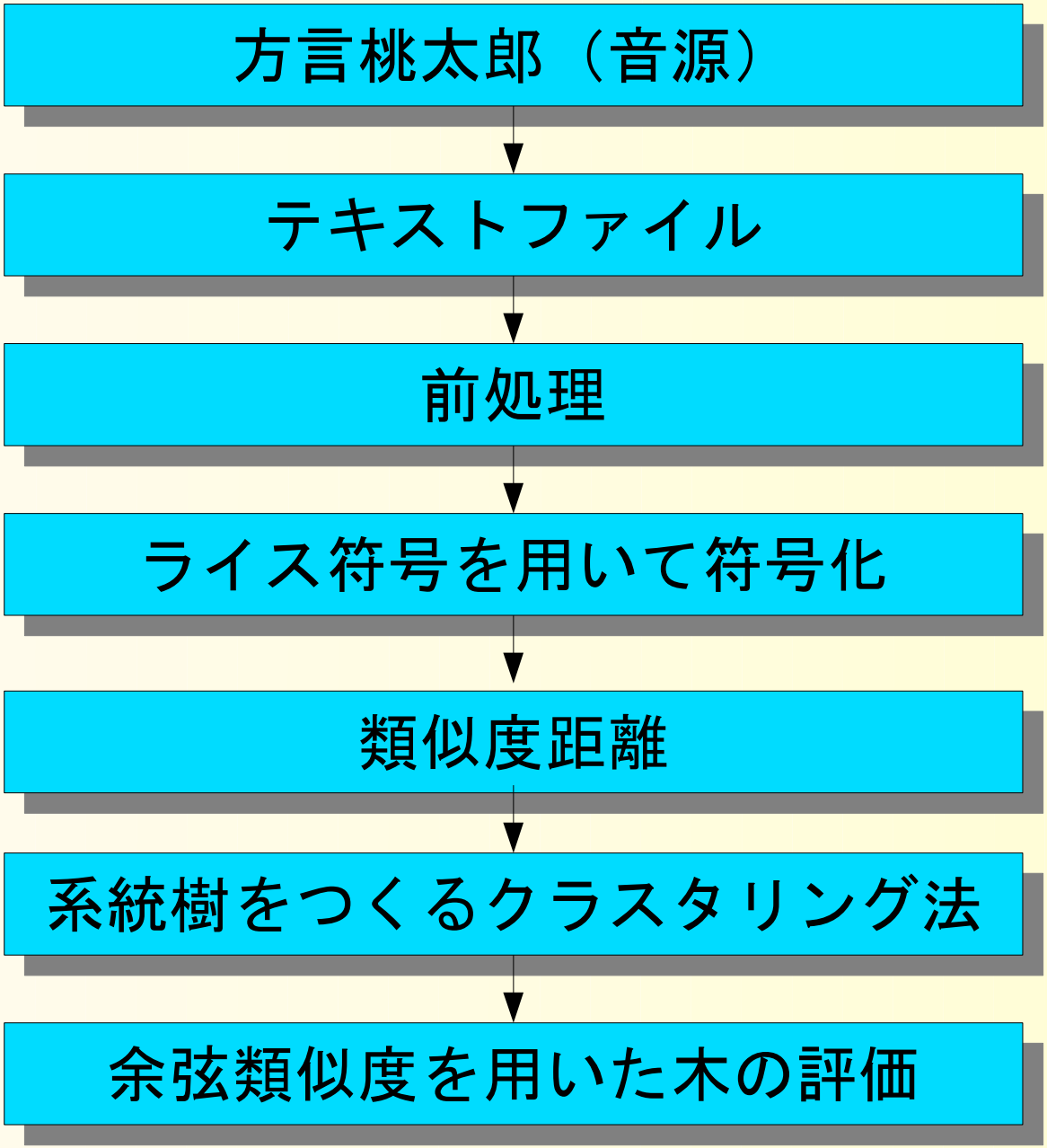
1. はじめに－目的

昨年の研究結果ではピッチの有用性を十分に確認出来なかった

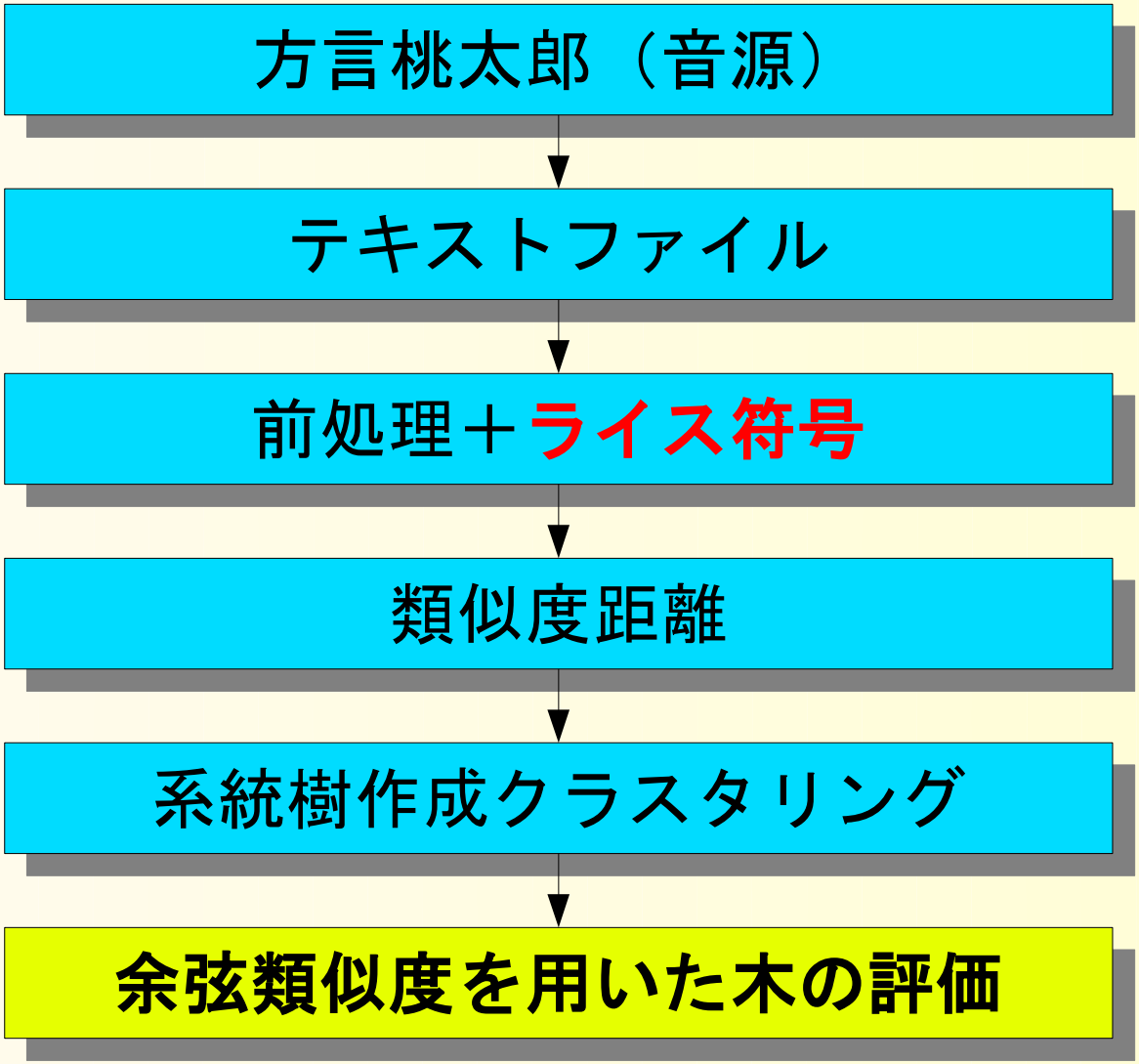


本研究では2008年研究のテキストファイルをもとに新たな研究項目を追加する

本研究の手順


方言桃太郎（音源）
テキストファイル
前処理
ライス符号を用いて符号化
類似度距離
系統樹をつくるクラスタリング法
余弦類似度を用いた木の評価

本研究の手順


方言桃太郎（音源）
テキストファイル
前処理＋ライス符号
類似度距離
系統樹作成クラスタリング
余弦類似度を用いた木の評価

研究動機

S(T) 値はQuartet Method以外では
評価値が同一になり正当に評価されているのか疑わしい

TV_1 差の平均を求めているが
平均では情報に大きなバラツキがあると
情報が失われてしまう