秘匿共通集合計算プロトコルを用いた就職活動支援システム“JHT”

研究指導 菊池浩明 教授
東海大学 電子情報学部 情報メディア学科

5ADM1123 香川 大介

目次

第1章 はじめに

1.1 背景

現在、就職活動において企業の情報を収集する方法は多くある。たとえば、新聞、就職支援サイトなどからの情報や、ネットワーク上の掲示板、SNS まで多彩にある。
その中で友達と協力し合い説明会や面接などの情報を交換しあえば上記の方法では得られない、貴重な情報を得られる可能性がる。そしてそれは就職活動をより有利に進める
相手の条件として、一つ目は自分が目指す企業の情報を持つ、二つ目は相手が虚偽をすることない信頼できる相手であることが必要である。一つ目は同系統の職種に就職するだろう同じ学部、学科の人間。二つ目として友人ある人間。このことから相手として同じ学部・学科の友人が適当である。そこで同じ学科の中で自分の志望企業のリストをホームページ上で公開すると共通の志望企業以外の志望企業も分かってしまう。自分の志望企業は友人に対しても基本的に隠したいし、またこの方法では友人以外に知られてしまう。

1.2 研究目的

そこで、本研究では、[1]の手法を応用し、それぞれの志望企業を隠したまま、共通の志望企業のみを抽出するシステム”Job Hunting Together(JHT)”を構築する。

第2章 要素技術

2.1 準同型暗号

準同型性を有する暗号方式。RSA暗号、ElGamal暗号など整数論ベースの公開鍵暗号方式でこの特徴を持っている。
この暗号は二つの暗号文 E[m1],E[m2]を与えられたとき平文や秘密鍵なしで

$$E[m_1]E[m_2] = E[m_1 + m_2]$$

$$E[m_1]^{m2} = E[m_1 m_2]$$

を行うことができる性質を持っている。

2.2 秘匿共通集合計算プロトコル[1]

集合を多項式で表し、準同型暗号を利用して、共有集合を求めるプロトコルである。
例えばクライアント{2,3}, サーバ{2,4}を持っているとする。クライアントはまず集合を多項式で表す

$$f(X) = (X-2)(X-3) = X^2 - 5X + 6$$

この多項式の係数を今回は変形ElGamal暗号で暗号化し, 次数とともにサーバに送信する

サーバは自身の持っている集合を代入しR(Randam)をかけ,代入した値を足す

$$f(2) = (2-2)(2-3) * R + 2$$

この作業を準同型暗号の性質を使い暗号文のまま行うためサーバは多項式の中身がわからない。
サーバはこの結果をクライアントに送る。
クライアントは多項式の暗号文を解凍し
準同型暗号を利用して、共有集合を求めるプロトコルである。

第3章 Job Hunting Together (JHT)

3.1 概要

二者間のユーザの一方が自分専用のサーバを運用している。それぞれクライアント、サーバの役割に分かれて、秘匿共通暗号計算プロトコルを用いて共通の志望企業を得る。

3.2 構成

・JHTC.java　クライアントプログラSwingアプリケーションとして作成。
サーバのアドレス、ポート、クライアントの志望企業リストの指定。
そしてサーバと通信し共通の志望企業リストを表示。

- mai()
  JHTの一連の関数を実行する
- ex()
  公開鍵が存在するか。ない場合は鍵を生成する
- hash()
  多項式のハッシュ関数を使った。多項式を複数の多項式に分け，次数の削除を実行する
- pg_read()
  公開鍵，秘密鍵を読み込む
- tushin()
  サーバと暗号化した文のやり取り，また共通な積をサーバに送信する
- angou()
  多項式の係数を暗号化する
- ans()
  変形ElGmal暗号の複号結果とマッチングするため，自身の持っている集合を元gを集合の分だけべき乗する
- hukugou()
  サーバより帰ってきた暗号文を複号する
- exp2()
  集合から式の展開を行う
- file_read()
  ファイルから企業リストを読み込む
- ad_file_name_read()
  アドレスとポート、企業ファイル名を読み込む
- ad_write()
  アドレスとポート名の初期設定を変える
- file_name_write()
  企業リストファイル名の初期設定を変える

・ JHTS.java　サーバプログラム。
Tera termやターミナル等のターミナルエミュレーターを用いて自身のサーバに接続し
サーバ上で起動するプログラム。
クライアントに対する応答、および共通の志望企業リストの表示。

- ex()
  公開鍵が存在するか。ない場合は鍵を生成する
- hash()
  多項式のハッシュ関数を使った。多項式を複数の多項式に分け，次数の削除を実行する
- ad_file_name_read()
  アドレスとポート、企業ファイル名を読み込む
- file_read2()
  ファイルから企業リストを読み込む
-mat()
  クライアントから返された共通企業リストのマッチング
-main()
  JHTSプログラムの一通りの流れ
-compute()
  秘匿共有集合プロトコルのサーバ部分

・ MkKey.java
公開鍵の生成
・ MKPrime.java
原始元gおよびp,qの生成

3.3 企業名フォーマット

志望企業リスト比較時に書式の違いによる共通企業の見逃しをなくす為フォーマットを設定する。

・ txt形式で一行に一社
・ 文字コードはSHIFT-JIS
・ 英字については半角、漢字、カタカナについては全角で記入

・ 株式会社を除く
株式会社 菊池研究室　→ 菊池研究室
東海University 株式会社 → 東海University

・ 略称がある場合は、略称で記入
KSS（香川システムズソリューション）　→ KSS

3.4 実行例

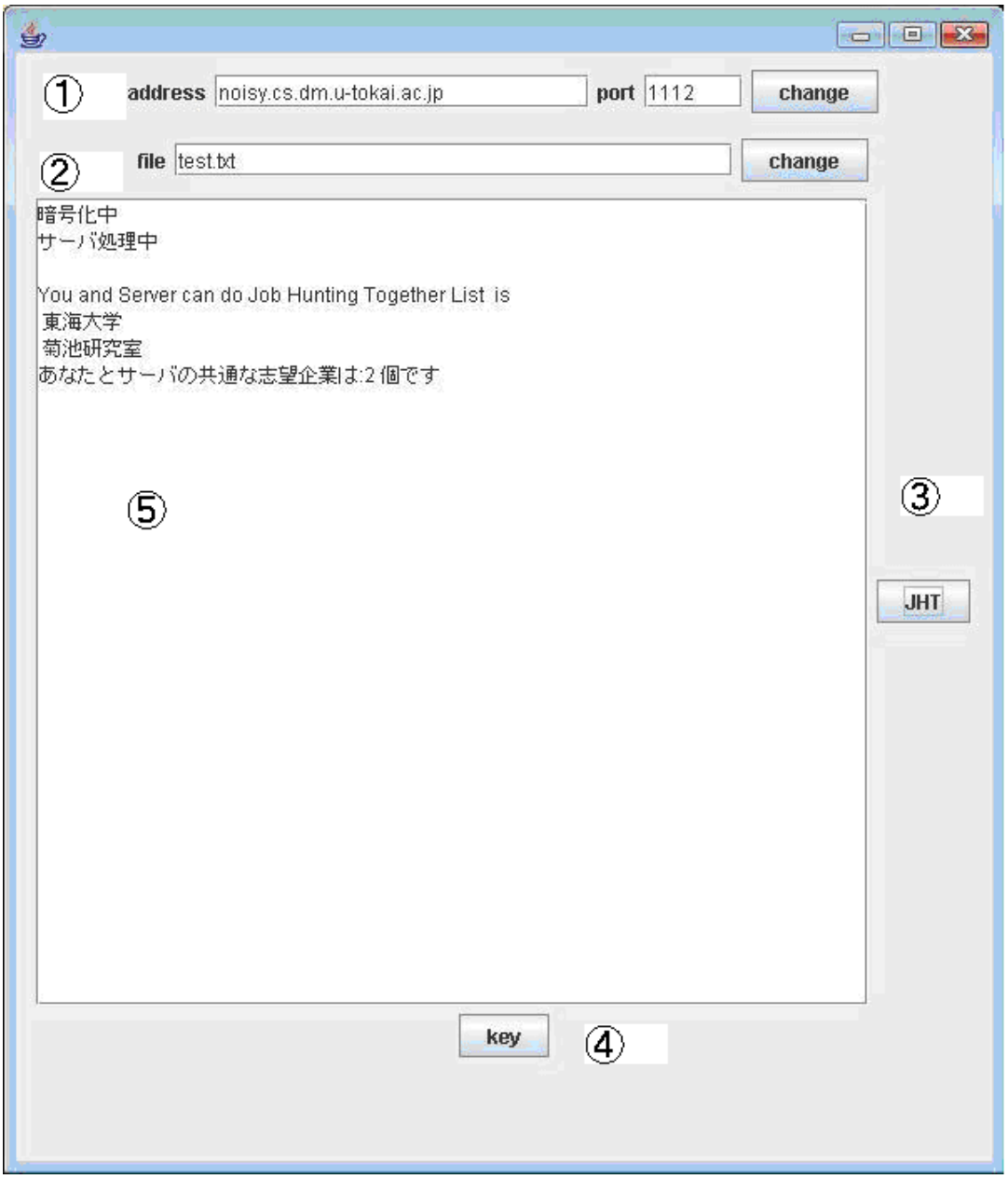
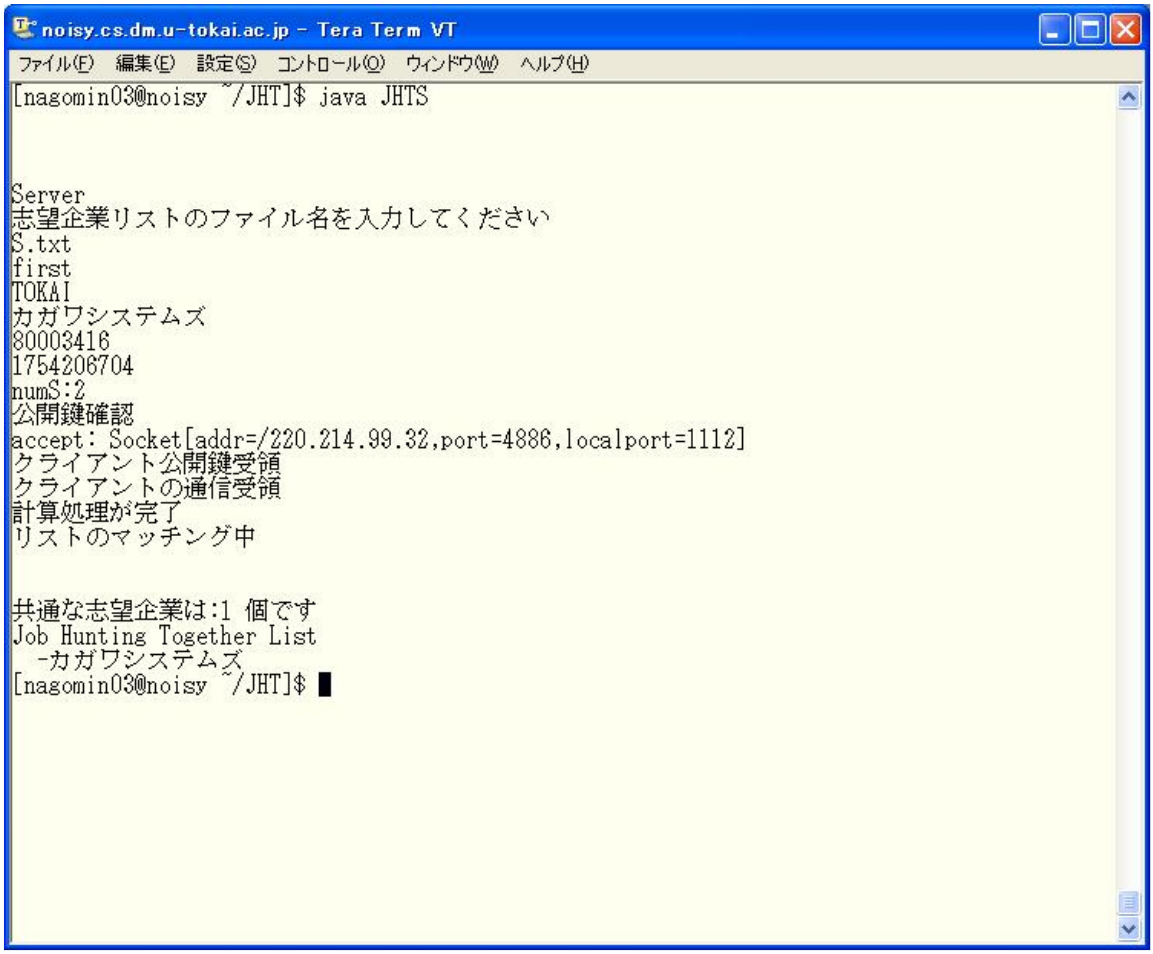


図1.1　クライアント表示画面

① サーバのアドレス、ポート番号を入力
② クライアントの志望企業リストのファイル名
③ サーバとのやり取りを開始
④ 秘匿共通集合計算プロトコルに使う公開鍵の生成
⑤ 現在の処理状況およびサーバとの共通志望企業の表示


noisy.cs.dm.u-tokai.ac.jp - Tera Term VT
ファイル(F) 編集(E) 設定(S) コントロール(O) ウィンドウ(W) ヘルプ(H)
[nagomin03@noisy ~/JHT]$ java JHTS
Server
志望企業リストのファイル名を入力してください
S.txt
first
TOKAI
カガワシステムズ
80003416
1754206704
numS:2
公開鍵確認
accept: Socket[addr=/220.214.99.32,port=4886,localport=1112]
クライアント公開鍵受領
クライアントの通信受領
計算処理が完了
リストのマッチング中
共通な志望企業は:1 個です
Job Hunting Together List
-カガワシステムズ
[nagomin03@noisy ~/JHT]$
図1.2　サーバ表示画面

第4章　実験

4.1　実験目的

　共通志望企業がどの程度出るのかの調査。

4.2　実験

　被験者にそれぞれ志望企業のリストを用意させ、互いに JHT を使い共通の志望企業を抽出させる。今回はその結果抽出された志望企業を報告してもらった。

4.3　実験結果

企業数は10社～30社　平均19.8社
共通な志望企業の平均は1.7社

表1.1　共通企業数

	A	B	C	D	E
A		0	3	3	1
B	0		0	1	1
C	3	0		4	1
D	3	1	4		3
E	1	1	1	3	

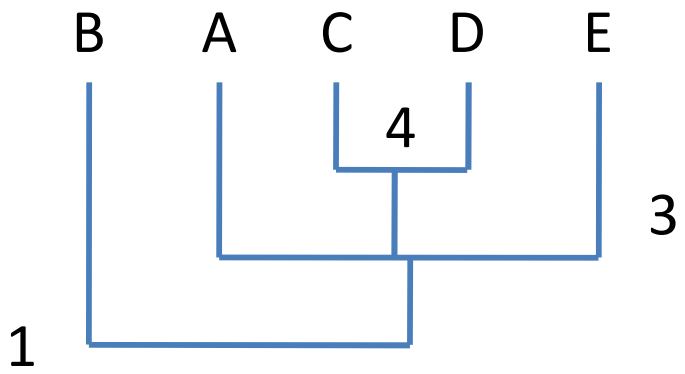


図1.3　最短距離法

第5章　結論

秘匿共通暗号計算プロトコルを用いることでお互いに志望企業リストを秘匿したまま共通な志望企業の抽出方法を構築した。また、実際に使用し共通の志望企業が抽出されること、それ以外の志望企業がわからないことを検証した。本実験では同じ研究室内で行なったが同じ学科で行なった場合など目指す職種が同じならば、JHT を使い共通の志望企業を抽出、その後企業についての情報交換、相談を行うという本システムが十分機能するだろうと考察できる。本方式は信頼できる友人同士で行うことを前提にしているので、今後の発展としては他人同士で使用できるような方法や二者間ではなく一対多で使用できる方法を考え出すなどがある。

謝辞

本研究を行うにあたり,暖かいご指導を受け賜りました東海大学情報通信学部通信ネットワーク工学科菊池浩明教授に深甚なる感謝を申し上げます。

また、システム構築など常に貴重な助言をして頂いた菊池研究室大学院生の磯崎邦隆氏に厚く感謝を申し上げます。最後に、本研究に協力して下さった菊池研究室および中西研究室、内田研究室の皆さまに感謝の意を述べると共に、謝辞とさせていただきます。

参考文献

[1] M.J.Freedman, K.Nissim, and B.Pinkas “Efficient Private Matching and Set Intersection”, EUROCRYPT pp.1-19, 2004

余禄　夏休み課題　PAM を使用したクラスタリング

1. 概要

中西、内田、菊池研究室の 3 研合同にて行う夏合宿は、他の研究室の人達と共に研究　　を発表する場である。そして研究に対する興味や質問は他人であるより友達であるほうがある。

　そこで趣味が合えば会話も弾み友達になりやすいという考えから趣味に関するアンケートを取り、それをクラスタリングにかけ同じ趣味のグループに分類、発表することによって友達になる手助けをするものである。

2. 実験手段

５０項目のアンケートに菊池研究室および他研究室 33 人に回答してもらい、
その回答をクラスタリング手法である PAM によって趣味、考えが近いグループに
分けていくものである。

[ネット]　SNS（mixi など）を利用している	Y　or　N

［ネット］ 2ch を利用する	Y or N
［ネット］ blog、ホームページを持っている	Y or N
［ネット］ オンラインゲームを行っている	Y or N
［ネット］ ニコニコ動画の「J の友人の動画」はすごいと思う	Y or N
［ネット］ 初音ミクを持っている	Y or N
［ネット］ Mac ユーザである	Y or N
［本］ ハリー・ポッターを読んだことがある	Y or N
［本］ シャーロックホームズを読んだことがある	Y or N
［本］ ズッコケ三人組みを読んだことがある	Y or N
［本］ チョコレートアンダーグラウンドを読んだことがある	Y or N
［本］ ジキルとハイドを読んだことがある	Y or N
［本］ ファーブル昆虫記を読んだことがある	Y or N
［アニメ］ エヴァンゲリオンが好きだ	Y or N
［アニメ］ 涼宮ハルヒが好きだ	Y or N
［アニメ］ 宇宙戦艦ヤマトが好きだ	Y or N
［アニメ］ 秒速5センチメートルが好きだ	Y or N
［アニメ］ クレイモアが好きだ	Y or N
［音楽］ ビートルズが好きだ	Y or N
［音楽］ イエペスの「禁じられた遊び」知っている	Y or N
［音楽］ エンヤが好きだ	Y or N
［音楽］ ブラックビスケッツが好きだ	Y or N
［音楽］ 水樹奈々が好きだ	Y or N
［音楽］ のまねこが好きだ	Y or N

表 1.1　アンケート項目の一部抜粋

3. PAM（中心点クラスタリング）

高速性、高次元のクラスタリングにおいてシンプルなアルゴリズムを目指して作られたクラスタリングである。方法としてはデータベースの任意のデータ A,B に対して、A,B に含まれる属性を比べていき、どれくらい異なっているかの相違度 $f(A,B)$ を用いる。

この手法ではこの相違度が距離である必要はない。ここでは k 個のデータ $O_1, O_2, ..., O_k$ をクラスタの代表データとして選び、対応するクラスタ $C_1, C_2, ..., C_k$ を作成する。

ここで、各データ A を $\min_{1\le j\le k} f(A,O_j)$ を達成する O_j、すなわちもっとも類似している代表のクラスタ C_j に分類する。代表データの選択に対しては最初適当に選択してきて、そ

れを逐次的に改良していくというアプローチをとる。ここでクラスタリングで対応しているデータと代表データの相違度の総和をポテンシャル（potential）（クラスタリングの評価関数）として選び、このポテンシャルが小さくなるように改良する。

数値で書くとポテンシャルは $\sum_{A \in S} \min_{1 \le j \le k} f(A, O_j)$ と書ける。

ここでは S はデータベースのレコード全体である。

具体的には、現在の代表データのうち一つ O_j を、別のデータ X にとりかえたとする。

このとき、C_j 以外の各クラスタ C_i では、その中のデータ A で $f(A,X) < f(A,O_i)$ であるものは、X クラスタに配属されなおす。このとき、ポテンシャルは $f(A,X) - f(A,O_j)$ だけ変化する（この変化は必ずポテンシャルを減少させる）。

クラスタ C_j の中のデータでは A は X のクラスタに配属されるか、もしくは今までの代表データで 2 番目に類似しているもののクラスタに配属される。この 2 番目に類似している代表データを $Second(A)$ と書くと、ポテンシャルは $\min\{f(A,Second(A)), f(A,X)\} - f(A,O_j)$ だけ変化する（この変化はポテンシャルを増加する可能性がある）。PAM では、ポテンシャルの増加が負である（すなわちポテンシャルが減る）変更を探して、変更を行う。もしこのような変化がなくなったら、その時点でクラスタリングを出力する。

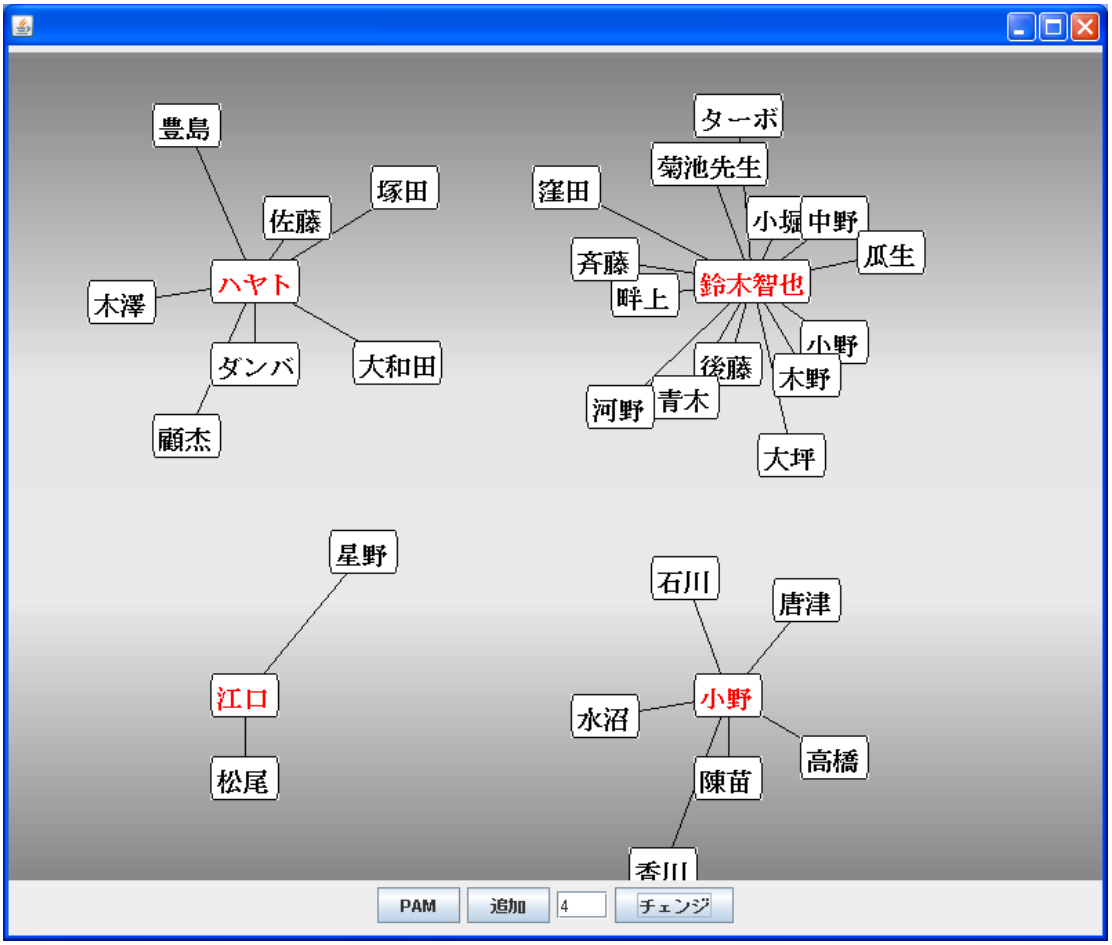
4. データフォーマット

今回使用するプログラムのデータフォーマットについて解説する。

Txt 形式のファイルに、一行につき一つのデータとする。データ内容に関しては
最初にデータを識別するための名前を 1 つ、その後コンマを挟みそのデータの持っている属性を整数でコンマを挟みながら配置する。そして改行コードをもってデータの入力を終了とする。

5. プログラム実行画面

図 1.1　プログラム実行画面



PAM　　：実行ボタン、データを読み込み PAM にデータをかけその結果を表示させる
追加　　：データを追加で読み込み PAM を再度かけなおす
チェンジ：代表となるデータの数（クラスタの数）を変更する
*相違度の大小によって 3 段階に距離が別れる

6.　実行結果

図 1.1 において右上の中心点に多くデータが集まっている理由はアンケートにおいて
ネット、本、アニメなどジャンルが別れているものに対しジャンルに偏りなく数個ずつ
回答されていることによって、ある同じ趣味の偏りを持つグループではなく平均的な回答をした
グループができてしまいそれぞれの趣味のグループに分けることに失敗している。

7.　追加実験

追加実験ではグループ分けが成功しなかった理由が PAM を使ったことにもあったので　はと考え、それぞれ同じような属性をもったグループ A,B,C のデータと、それに属さない D のデータ（表 1.2）を作成し PAM の実験を行った。その結果、代表点 k =3 の時は A,B,C
それぞれのグループにまとめられ D は C のグループに所属していたが、k ＝4 にした場合、期

待していた A,B,C,D のグループに分かれるのではなく図 1.3 のようになった。
k=3 の時は、データ B2 を中心に B グループを作成しいくつかの同じ属性の重なり元にグループを作成し PAM として順調な結果を出した。しかし k=4 のときは図 1.3 では隠れているが B1 と B2 を入れ替え B2 を代表点とし,同時に B1 と D1 を入れ替え D1 を代表店とすればポテンシャルが 10→6 に減りよいグループわけができるはずが、プログラムを設計した際代表点を 1 つずつ更新していく使用のため 2 つ同時に代表点の変更ができないためにこの結果を招いたと思われる。

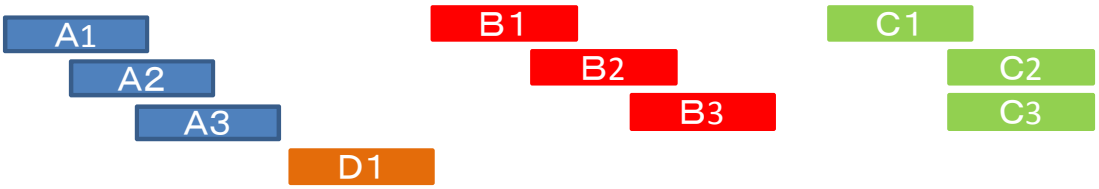


図 1.2　追加実験のデータについて

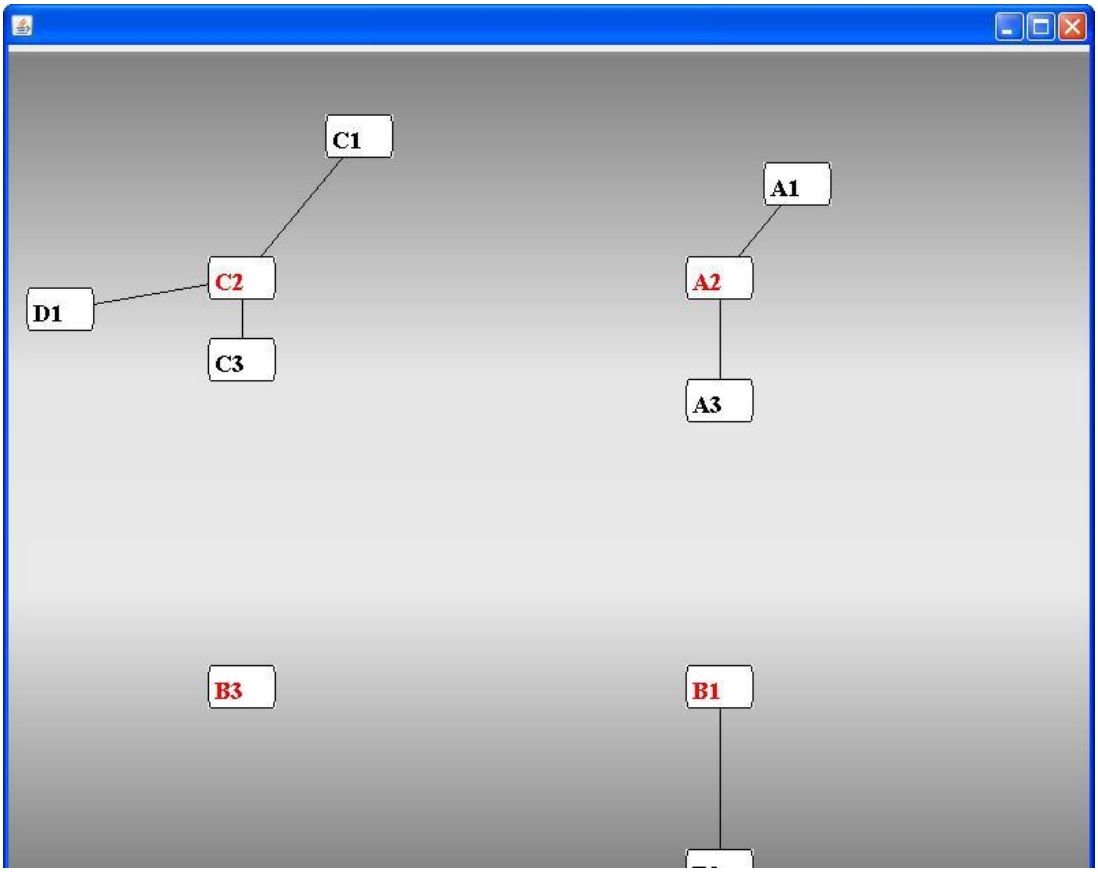


図 1.3 追加実験結果

表 1.2　A-D の相違度表

	A1	A2	A3	B1	B2	B3	C1	C2	C3	D1
A1		2	4	10	10	10	10	10	10	10
A2	2		2	10	10	10	10	10	10	10
A3	4	2		10	10	10	10	10	10	10
B1	10	10	10		6	10	10	10	10	10
B2	10	10	10	6		6	10	10	10	10
B3	10	10	10	10	6		10	10	10	10
C1	10	10	10	10	10	10		8	8	10
C2	10	10	10	10	10	10	8		0	10
C3	10	10	10	10	10	10	8	0		10
D1	10	10	10	10	10	10	10	10	10	

8.　考察

　失敗した理由にアンケートの内容に偏りが少なかったこと、それによって中心点データに平均的なデータを選ぶ PAM の手法が悪い方向に働いてしまいグループ分けが成功しなかった。と同時にプログラム的な不備によりグループ分けに多少の不備が働いた可能性があるが研究室ごとや 3 研全体のグループ分けによる話題の提供はなったと考える。

9.　参考文献

[1] 福田剛志, 森本康彦, 徳山豪, “データサイエンス・シリーズ③　データマイニング”,　共立出版, 2001.