Vol.2019-DBS-169 No.5
Vol.2019-IFAT-136 No.5
2019/9/10

Twitterの反応を用いたニュース全体像の理解支援のための可視化手法

池田　将[1,a)]　牛尼　剛聡[2]

概要：近年、Twitter などの SNS を利用してニュースを知り、SNS に投稿された反応を読むことでニュースの理解を深めるユーザが増えている。一方、SNS でニュースや反応を閲覧する際に、ユーザのフォローネットワークに基づき選択的に情報が提示されるため、エコーチェンバーと呼ばれる情報の偏りが起こることが問題となっている。そこで本稿では、Twitter の反応を利用しニュースの全体像の理解支援を行うための可視化手法を提案する。具体的には、Twitter で投稿されたニュースに対する反応としてリプライ、引用リツイートを用い、ニュース自体の特徴語と反応の特徴語を抽出する。抽出した特徴語を利用して、ニュースや反応の特徴および他のニュースとの関連性をわかりやすく可視化する。

1. はじめに

SNS(ソーシャル・ネットワーク・サービス) とは、他のユーザとの交友関係に基づいてコミュニケーションを行うインターネット上のサービスである。代表的な SNS には、ツイートと呼ばれる短い文章を投稿し、他のユーザがそれを閲覧、返信をすることで、コミュニケーションを行う Twitter や、自分のアカウントを作成することで、知り合いや共通する趣味を持つ人同士で交流することができる Facebook、画像や短い動画を共有する Instagram などがある。それぞれのサービスの国内の利用者は、Twitter が約 4,500 万人 (2018 年 10 月時点)、Facebook が約 2,800 万人 (2019 年 1 月時点)、Instagram が約 3,300 万人 (2019 年 3 月時点) いるとされる [1]。近年では、多くの企業や公的機関が SNS 上にアカウントを作成し、情報を提供している。

新聞社やテレビ局、インターネットのニュースサイトなど、多くのメディアが Twitter 上にアカウントを開設し、ニュースの見出しを投稿している。ユーザは興味のあるニュースがあった場合、そのツイートから参照される Web ページの記事を読むことができる。また、ニュースのツイートに対して行われたリプライや引用リツイートなどといった反応を読むことができる。ニュースに対する反応の中には、ニュースを理解する際の様々な観点や、関連するニュースについて言及されていることがあり、これらのニュースの観点や、他の関連するニュースを知ることで、ニュースの全体像を理解することができる。

SNS でのニュース閲覧と、既存のマスメディアでのニュース閲覧と異なる点を以下に示す。新聞やテレビの場合、ニュースが体系的にまとめられているが、SNS ではユーザが気になったニュースを選択的に閲覧する。また、マスメディアでは記者や解説員、コメンテーターのような人々がユーザのニュース理解を促進するために役割を果たしているが、SNS でのニュース閲覧においては、そのニュースに詳しい一般ユーザがその役割を果たしている。

SNS でニュースを理解する際、いくつかの問題点が存在する。ニュースはタイムライン上に断片的に現れる。そのため、全てのニュースを閲覧することは難しく、必要な情報を見逃してしまうことがある。また、日常的に SNS を利用しないユーザには、既存のブラウジング手法ではニュースの全体像を理解するのは難しい。さらには、ユーザのフォローネットワークによって選択的に情報が提示されるフィルターバブルのような問題も存在する。

そこで、本研究では、ユーザがニュースやその論点について正しく理解するための可視化手法を開発することを目的とする。具体的には、SNS 上のニュースとその反応を利用してニュース間の関連を可視化するインターフェースの

[1] 九州大学大学院芸術工学府
4-9-1 Shiobaru, Minamiku, Fukuoka, 815–8540, Japan
[2] 九州大学大学院芸術工学研究院
4-9-1 Shiobaru, Minamiku, Fukuoka, 815–8540, Japan
[a)] ikeda.sho.294@s.kyushu-u.ac.jp

Vol.2019-DBS-169 No.5
Vol.2019-IFAT-136 No.5
2019/9/10

開発を目的とする。本稿で提案するインターフェースでは、ユーザに対し、閲覧しているニュースの反応の中で特徴的に現れる語 (反応の特徴語) と関連する別のニュースを提示する。ユーザは反応の特徴語を知ることで、そのニュースにおいて注目されている観点を知ることができる。また関連するニュースを知り、閲覧しているニュースと比較することでそのニュースの位置付けを理解することができる。観点と位置付けを知ることでそのニュースの全体像を理解することができると考えられる。

本研究ではニュース間の関連を可視化するインターフェースを実現するために主に 2 つのプロセスを行う。まずニュース本文の特徴の類似度をもとに、同じニュースのクラスタリングを行う。次に、ニュースの反応の特徴を元に関連するニュースを繋げる。最終的にユーザにはニュースの反応に現れる特徴語と関連するニュースを提示する。

本稿ではこれらの手法についての具体的な手順、および行った実験についてその結果と考察について述べる。第 2 章では、本稿と関連する研究について、第 3 章では、本稿で提案するシステムの概要について、第 4 章では提案手法について、第 5 章では実験について、第 6 章では今後の課題についてそれぞれ述べる。

2. 関連研究

本研究では、SNS 上のニュースとその反応を利用してニュース間の関連を可視化するインターフェースの開発を目的としている。本章では、関連研究として、ニュース閲覧に関する研究と SNS とニュースに関する研究について述べる。

2.1 ニュース閲覧に関する研究

Liu ら [2] はジャーナリストは自身が属する社会集団の文化的規範と価値に影響を受けニュースを制作することから、読者に異なる社会集団から発信された記事をまとめて提示することで多様な視点を提供する LocalSavvy というパラダイムを提案している。

神嶌ら [3] は情報推薦システムにおけるフィルターバブル問題を解決するため、ユーザが選んだある特定の視点・情報に対する中立性を保証する情報推薦システムを提案している。

片岡ら [4] は Google 検索においてユーザが気付かないうちに閲覧する情報に偏りが生じる問題を解決するため、ユーザにパーソナライズされた検索結果とそうでない結果との差をフィルターバブルの度合いとしてユーザに提示し、認知させることで、ユーザにフィルターバブルの存在を自覚させ、ユーザが情報の探索領域を広げることを促すユーザ・インターフェースを提案している。

切通ら [6] はユーザがニュースイベントに対して様々な視点から理解するのを支援するためのニュースアプリケーション NewsSalad を提案している。NewsSalad では、ニュースを意見、視点、詳しさの 3 つの尺度で定量化し、ユーザに対し閲覧中の記事と同じイベントでかつその差異の尺度が最も大きい記事を提示する。

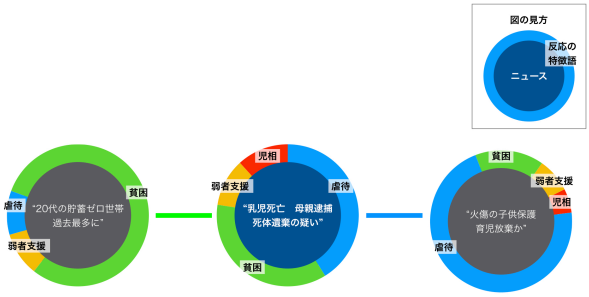

図 1 可視化の例

本研究は、これら既存の手法と異なり、SNS を利用してニュースに関する多様な観点やそのニュースの位置付けをユーザに知らせることにより、ニュースに対するより深い理解を促すことを目的としている。

2.2 SNS とニュースに関する研究

川口ら [7] はユーザが所属する SNS のコミュニティに偏りがあるために、SNS 上で閲覧するニュースや反応が偏り、ユーザの中立的観点が損なわれるという問題を解決するため、ニュースに対する日常的関心度に基づく反応ユーザの分布の傾向を示す指標として「ポピュラリティ」を提案し、ニュースや反応の中立性を推定する手法を提案している。

また、本稿ではニュースを閲覧する際に SNS の反応の情報を利用しているが、SNS での感情の分析を行う研究として、Jiang ら [8] のツイートのターゲットに対する特徴を利用した主観分類、極性分類を行う手法がある。しかし、SNS において反応を利用してニュースを特徴づける研究はあまり行われておらず、その点で本研究について新規性がある。

3. システムの概要

第 1 章でも述べたとおり、本研究ではユーザがニュースやその論点について正しく理解することを目的としている。そのため、SNS を利用して、ニュースの全体像を簡単に理解できる可視化手法を開発する。この可視化手法では、普段 SNS を利用しないユーザや、多くの投稿を閲覧する時間がないユーザが手軽にニュースとその論点を理解でき、ニュースの客観的な位置付けを知ることができるようにする。開発した可視化手法を利用して SNS 上のニュースとその反応を利用してニュース間の関連を把握可能なインターフェースの開発を行う。本研究で提案するインターフェースの例を図 1 に示す。

図 2　ニュースツイートと反応の例

本手法のインターフェースではユーザに対し、ニュース、ニュースに対する反応の中に特徴的に現れた単語 (反応の特徴語)、ニュース間の関連を提示する。ニュースは円で表示され、同じ内容のニュース記事がまとめられている。反応の特徴語はニュースの外側の円として表される。ニュース同士を結ぶ線は、それらに関連があることを示している。また、ユーザがこれらをタップすることで、ニュース本文や、反応の特徴語が含まれる反応ツイートを閲覧できるようにする。

本手法を実現するための処理として、同じニュースのクラスタリング、反応の特徴の抽出、反応の特徴に基づいた関連づけを行う必要がある。次章では、ニュースの特徴語を抽出する手法やその特徴ベクトルを作成する具体的な方法について述べる。

4. 提案手法

本手法の大まかな流れは、データの取得、ニュースのクラスタリング、反応の特徴に基づいた関連付けである。本章ではそれを実現するために必要な処理について述べる。

4.1 データの取得

本手法では、Twitter に投稿されたニュースのツイート、そのニュースに対してなされた反応のツイート、Web サイト上にアップロードされているニュース記事の本文を利用する。本手法で扱うデータの例を図 2 に示す。このうち、反応のツイートとは、あるニュースについてのリプライ、引用リツイートとする。

本手法において扱うデータのうち、ニュースのツイート、反応のツイートは TwitterAPI によって収集する。ニュースのツイートは、ニュースを配信しているアカウントを指定し、そのアカウントが配信しているツイートを取得する。反応ツイートは、ニュースを配信しているアカウントのスクリーンネームをクエリとして検索し、リプライと引用リツイートを収集する。収集されたデータの中に、どのツイートに対しての反応かというデータが含まれている。ニュース記事の本文は、収集されたニュースのツイートの情報に含まれる Web ページの URL を指定し、スクレイピングを行うことで収集する。

収集したデータについて、以下のように前処理を行う。まず、反応のツイートについて、リプライにつけられている返信先や、引用リツイートにつけられている URL など、ツイート中の不要な情報を消去する。次に、反応のツイートとニュース記事の本文について、文章を単語ごとに分割するため、形態素解析を行う。形態素解析には、オープンソースの日本語形態素解析エンジンである MeCab[9] を利用する。

4.2 特徴語の抽出

本手法では、ニュースの本文や、ニュースの反応の中に特徴的に現れる単語を利用して、ニュースの特徴付けを行う。そのために、文書中の単語の重要度を測る tf-idf 法を利用して、各単語の重要度を算出する。tf-idf とは、Jones[10] によって提唱された文書中の単語の重み付け指標である。tf-idf は単語の出現頻度 tf と、逆文書頻度 idf という 2 つの指標に基づいて計算される。$|D|$ を総文書数、$n_{i,j}$ を文書 d_j における単語 t_i の出現回数、$\sum_k n_{k,j}$ を文書 d_j におけるすべての単語の出現回数の和、$|\{d : d \ni t_i\}|$ を単語 t_i を含む文書数とすると、$tf_{i,j}$、idf_i、$tfidf_{i,j}$ は以下のように計算できる。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

$$idf_i = \log \frac{|D|}{|\{d : d \ni t_i\}|} \tag{2}$$

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i \tag{3}$$

tf は文書中に多く現れる単語の値が高くなり、idf は多くの文書に現れる単語の値が低くなる。これにより、全ての文書と比較して、その文書において特徴的に現れる語が何か求めることができる。

本研究では、ニュース本文や反応の中に含まれる名詞の単語に対して、tf-idf 値を求めることで、語の重要度を求める。ニュース本文に対しては、1 記事を 1 文書として、それぞれの語の tf-idf 値を求める。ニュースに対する反応については、1 つのニュースに対する全ての反応を結合し、それを 1 文書として、語の tf-idf 値を求める。

4.3 ニュースの特徴量の算出

ニュースのクラスタリングや、ニュースどうしの関連付けを行うために、ニュース本文や反応を特徴ベクトルにする必要がある。本手法では、Word2Vec 法を利用して特徴ベクトルを作成する。

Word2Vec[11] は、ニューラルネットワークを利用して単語の分散表現を得るシステムである。文章中のある単語

とその周辺にある単語を入力し、その単語の次にどの単語が出現するかという確率を出力するモデルについて、学習を繰り返すことで、中間層のパラメータを調整する。生成されたベクトルについて、単語を入力した時の中間層の値がその単語の分散表現を表す特徴ベクトルとなる。

本手法では、以下の手順でニュースの特徴ベクトルを算出する。まず、ニュース本文中にある全ての名詞の単語について、それぞれ Word2Vec を利用して特徴ベクトルを作成する。ニュースの特徴ベクトルはこれらの単語の特徴ベクトルの重み付き平均とする。ここでの重みは先に求めた単語の tf-idf 値である。あるニュース i のニュース本文の特徴ベクトル $\boldsymbol{nv}_i$ は、ニュース i のニュース本文に含まれる単語のベクトルを $\boldsymbol{wv}_{i,k}$、重みを $tfidf_{i,k}$ とすると、以下のように算出することができる。

$$\boldsymbol{nv}_i = \frac{\sum_k \boldsymbol{wv}_{i,k} \cdot tfidf_{i,k}}{\sum_k tfidf_{i,k}} \tag{4}$$

4.4 反応の特徴量の算出

本研究では、ユーザが閲覧するニュースの位置付けを知るために、ニュース同士の関連付けを行う。SNS におけるニュースに対する反応の中には、そのニュースに対する直感的な感想の他に、問題となっている点や原因、関連する他のニュースに関する言及がなされている場合がある。そこで、本手法では、反応の特徴量を元に関連付けを行う。

反応の場合もニュース本文の場合と同様に、Word2Vec を利用して特徴ベクトルを作成する。あるニュースに対する反応に現れる名詞の単語のうち、ニュース本文に現れる語を除いたものについて、それぞれ Word2Vec を利用して特徴ベクトルを作成する。反応の特徴ベクトルはこれらの単語の特徴ベクトルの重み付き平均とする。ここでの重みは先に求めた単語の tf-idf 値である。あるニュース i に対する反応の特徴ベクトル $\boldsymbol{mv}_i$ は、単語のベクトルを $\boldsymbol{wv}_{i,k}$、重みを $tfidf_{i,k}$ とすると、以下のように算出することができる。

$$\boldsymbol{mv}_i = \frac{\sum_k \boldsymbol{wv}_{i,k} \cdot tfidf_{i,k}}{\sum_k tfidf_{i,k}} \tag{5}$$

5. 実験

本研究において、提案するインターフェースを実現するために、前章で述べたとおり、ニュース本文と反応の中で特徴的に現れる語を抽出する必要がある。そこで、ニュース本文の特徴語と、ニュースに対する反応の特徴語の抽出を行った。

5.1 データの準備

まず、本研究において扱うデータの収集を行う。ニュースは Yahoo!ニュース (@YahooNewsTopics)[12] で配信されているニュースを対象として、TwitterAPI を利用しニュースツイートの収集を行った。Yahoo!ニュースを対象とした理由として、新聞社、通信社など様々なメディアからニュースを提供され、それを配信しているため、多様な視点からのニュースが配信されるとともに、集められる反応にも主張の偏りがないと考えたためである。次に Yahoo!ニュースに対して行われたリプライ、引用リツイートを収集した。本稿ではこれらを反応ツイートと呼ぶ。また、ニュース記事の本文は、Tweet につけられている URL を元にスクレイピングを行い収集した。本研究では、反応を利用してニュースの特徴をはかるため、そのニュースに対する反応がある程度必要である。そこで今回は、リプライと引用リツイートが 50 回以上なされているニュースを対象とした。結果として、2019 年 6 月から 7 月までの間に、これらの条件を満たす記事を 649 本収集することができた。

本研究では以下のような前処理を行なった。まず、反応のツイートについて、リプライにつけられている返信先や、引用リツイートにつけられている URL など、ツイート中の不要な情報を消去した。次に、反応のツイートとニュース記事の本文について、文章を単語ごとに分割するため、MeCab を利用して形態素解析を行った。各単語について、(ニュースまたは反応の Twitter ID, 語順, 単語, 単語の原型, 品詞) という形でデータベースに格納した。

表 1　ニュース本文の特徴語と tf-idf 値

ニュース	特徴語	tf-idf 値
【自民 2 千万円報告書もうない】	国対	0.378
自民党の森山裕国対委員長は、	森山	0.338
老後資金として公的年金以外に	書	0.248
「30 年間で約 2 千万円が必要」	年金	0.220
などとした金融庁の報告書を巡り...	報告	0.210
【サウジ 船攻撃でイランを非難】	イラン	0.358
サウジアラビアのムハンマド・	皇太子	0.284
ビン・サルマン皇太子は、	攻撃	0.250
アラブ紙が 16 日に掲載した	タンカー	0.250
インタビューで、中東で 13 日に...	隻	0.239
【育休取らぬ日本男性 国連指摘】	ユニセフ	0.331
国連児童基金は日本など	取得	0.300
41 カ国の政府による 2016 年時点の	休暇	0.260
子育て支援策に関する報告書を	41	0.208
発表し、日本男性の育休について...	育児	0.193

5.2 ニュースの特徴語の抽出

ニュース本文に現れた単語について、tf-idf 値を算出した。その結果の例を表 1 に表す。各ニュースのニュース本文に現れた名詞の単語について、tf-idf 値が高い順に特徴語を 5 語表している。

5.3 反応の特徴語の抽出

ニュースに対する反応に現れた単語について、tf-idf 値を算出した。その結果の例を表 2 に表す。各ニュースの反

表 2　ニュースに対する反応の特徴語と tf-idf 値

ニュース	特徴語	tf-idf 値
【自民 2 千万円報告書もうない】	隠蔽	0.156
自民党の森山裕国対委員長は、	選挙	0.126
老後資金として公的年金以外に	隠滅	0.122
「30 年間で約 2 千万円が必要」	文書	0.106
などとした金融庁の報告書を巡り...	都合	0.104
【サウジ 船攻撃でイランを非難】	黒幕	0.192
サウジアラビアのムハンマド・	ジャーナリスト	0.109
ビン・サルマン皇太子は、	アメリカ	0.090
アラブ紙が 16 日に掲載した	察し	0.065
インタビューで、中東で 13 日に...	根拠	0.061
【育休取らぬ日本男性 国連指摘】	カネカ	0.210
国連児童基金は日本など	転勤	0.082
41 カ国の政府による 2016 年時点の	会社	0.057
子育て支援策に関する報告書を	有給	0.052
発表し、日本男性の育休について...	男	0.040

応に現れた語のうちニュース本文に現れていない名詞の単語について、tf-idf 値が高い順に特徴語を 5 語表している。

5.4 考察

今回の実験ではニュース本文に現れる特徴語と、ニュースに対する反応に現れる特徴語を抽出した。ニュースに対する反応の特徴語には、例えば「【育休取らぬ日本男性 国連指摘】...」というニュースでは、「カネカ」、「転勤」のような語が特徴語として現れている。これは、2019 年 6 月の初頭に株式会社カネカにおいて、育児休暇を取得した男性社員が育児休暇明けすぐに転勤を命じられたというニュースが話題になっており、それらについて言及した反応が多くあった。このことはニュース本文では言及されておらず、これらの特徴語がニュースの関連付けに役に立つ可能性を示している。

6. 今後の課題

本稿では、ニュース本文とニュースに対する反応に現れる特徴語の抽出を行なった。今後はこれらの特徴を利用して、ニュースのクラスタリングやニュースどうしの関連付けを行なっていきたいと考えている。

ニュースのクラスタリングは、ニュース本文の特徴ベクトル $\boldsymbol{nv}$ を利用して、DBSCAN[13] で行うことを考えている。DBSCAN は密度準拠のクラスタリング手法であり、半径、密度をあらかじめ設定しておき、ある一点から決められた半径の中で密度を満たしていればクラスタを形成し、それを繰り返すことでクラスタを成長させていく手法である。クラスタ間の距離の閾値 Eps とクラスタを構成する最小のデータ数 $MinPts$ の 2 つを持ち、ある点 x から距離 Eps 以内にある点集合を近傍 $N_{Eps}(x)$ と定義し、以下の接続関係を満たす時、同じクラスタに分類する。

$$y \in N_{Eps}(x) \tag{6}$$

$$|N_{Eps}(x)| \geq MinPts \tag{7}$$

また、できたニュースのクラスタ nc_c の反応の特徴ベクトル cmv_c は、それぞれのニュース $\{n : n \in nc_c\}$ の反応ベクトルの平均して、以下のように求める。

$$\boldsymbol{cmv}_c = \frac{\sum_{\{n:n\in nc_c\}} \boldsymbol{mv}_n}{|nc_c|} \tag{8}$$

ニュースどうしの関連付けは、ニュースクラスタの類似度として求めることを考えている。ユーザが閲覧しているニュース nc_c とあるニュース nc_x の類似度を以下のようにコサイン類似度として求める。

$$NewsRelation(c, x) = \frac{\boldsymbol{cmv}_c \cdot \boldsymbol{cmv}_x}{|\boldsymbol{cmv}_c||\boldsymbol{cmv}_x|} \tag{9}$$

$NewsRelation$ の値が高いニュースどうしを関連のあるニュースとして提示する。

7. 終わりに

本稿では、SNS でニュースを閲覧する際に、従来のブラウジング手法では、そのニュースの論点が整理されていないという問題や、日常的に SNS を利用していないユーザがニュースの全体像を理解しにくいという問題を解決し、SNS を利用して誰でも簡単にニュースの全体像を理解できるようにするために、SNS の反応を利用してニュースどうしの関連を可視化する手法を提案した。また、自動的な可視化を実現するために、ニュースの記事本文とニュースに対する反応に特徴的に現れる語の抽出を行なった。今後は、これらの特徴を利用し、ニュースのクラスタリング、ニュースどうしの関連付けを行う予定である。

参考文献

[1] “【2019 年 7 月更新】主要ソーシャルメディアのユーザー数まとめ”,uniad,2019 年 7 月 3 日更新,https://www.uniad.co.jp/260204, 2019 年 8 月 5 日閲覧
[2] Jianhui Liu, Larry Birnbaum, ‘‘LocalSavvy : Aggregating Local Points of View about News Issues”,Proc. 1st International Workshop on Location and the Web, 2008, 33-40, 2008
[3] 神嶌 敏弘, 赤穂 昭太郎, 麻生 英樹, 佐久間 淳,“情報中立推薦システム”, 人工知能学会全国大会論文集 26,2012
[4] 片岡 雅裕, 橋山 智訓, 田野 俊一,“情報推薦システムにおいて閲覧する情報の偏りを気付かせる UI の設計”,31st Fuzzy System Symposium,2015.9
[5] 青木 伸也, 湯本 高行, 角谷 和俊, 新居 学, 高橋 豊,“論点に対する極性に注目したニュース記事からの編集意図の抽出方法”,Vol.2009-DBS-149 No.16,2009.11.21
[6] 切通 恵介, 楠見 孝, 堀江伸太朗, 馬 強,“多様性指向のニュースアプリの開発とその有用性評価”,DEIM Forum 2016,2016.3
[7] 川口 天佑, 牛尼 剛聡,“ポピュラリティ推定に基づいた SNS におけるニュースの中立的な理解支援”,DEIM Forum 2018,2018.3
[8] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, Tiejun Zhao,‘‘Target-dependent Twitter Sentiment Classification”, Proceedings of the 49th Annual Meeting of the

Association for Computational Linguistics, pages 151 - 160, 2011
[9] 形態素解析エンジン MeCab,http://taku910.github.io/mecab/
[10] Karen Spärck Jones, “A statistical interpretation of term specificity and its application in retrieval”, Journal of Documentation, 28, 11-21,1972
[11] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean,“Efficient Estimation of Word Representations in Vector Space”,Proceedings of the International Conference on Learning Representations,2013
[12] Yahoo!ニュース, https://news.yahoo.co.jp/
[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu,“A Density-Based Algorithm for Discovering Clusters”, KDD-96, 1996