Activating Participants Through Social Networks and Gamification in Undertourism Areas

(Study materials)

Jiří Kysela and Pavla Štorková (eds.)

University of Pardubice

Faculty of Electrical Engineering and Informatics

Activating Participants Through Social Networks and Gamification in Undertourism Areas

(Study materials)

Jiří Kysela and Pavla Štorková (eds.)

This publication is supported by the EEA grants 2014-2021 CZ-EDUCATION of the Project Activating Participants Through Social Networks and Gamification in Undertourism Areas. Number of project EHP-CZ-ICP-3-003.

Univerzita Pardubice

2023

Content

Introduction …………………………………………………………………………. 4

Jiří Kysela
E-Tourism …………………………………………………………………………. 5

Jiří Kysela and Pavla Štorková
Augmented Reality in E-Tourism ……………………………………………………… 20

Jan Panuš
Gamification …………………………………………………………………...... 31

Jerry Chun-Wei Lin
A Survey of Social Network Analysis ……………………………………………….….. 49

Jerry Chun-Wei Lin
AI-Empowered Decomposition and Evolutionary
Computation for the Closed Pattern Discovery …………………………….……… 73

Usman Ahmed and Jerry Chun-Wei Lin
Mining of High Expected Utility Patterns from the Uncertain Databases …...………… 97

Introduction

The study materials deal with interdisciplinary topics in the fields of tourism, history and information and communication technologies. There are six chapters in this book. It interconnects these disciplines into a system and explores and realises the possibilities of applied informatics, using new technologies for the innovation of heritage tourism. As a result, it brings enrichment to the pillars of modern e-tourism/m-tourism, not only theoretical within the chapters of the study materials, but also in the form of a practical output of a mobile web application, in which the authors of these chapters participated, transforming theoretical information into practice. The application, which is available on the Nezmizelo.cz website, allows to visualise the historically accurate images of long-vanished monuments in the Czech towns of Pardubice and Litomyšl in the form of 3D models displayed to tourism participants thanks to augmented reality directly in the places of their original occurrence. The activation of tourism participants is also achieved through the connection of the application with the geosocial network (Facebook) where the principles of gamification are used, i.e., a technique that seeks to increase the interest of users using gaming principles in the media industry.

The proposed theoretical system based on tourism in connection with information and mobile communication technologies was thus implemented in the form of a locally contextual application with augmented reality. The application is freely available to tourism participants in the mentioned locations, where it provides local information in 3D graphics based on the visitor's location in the tourist destination.

E -Tourism

Jiří Kysela

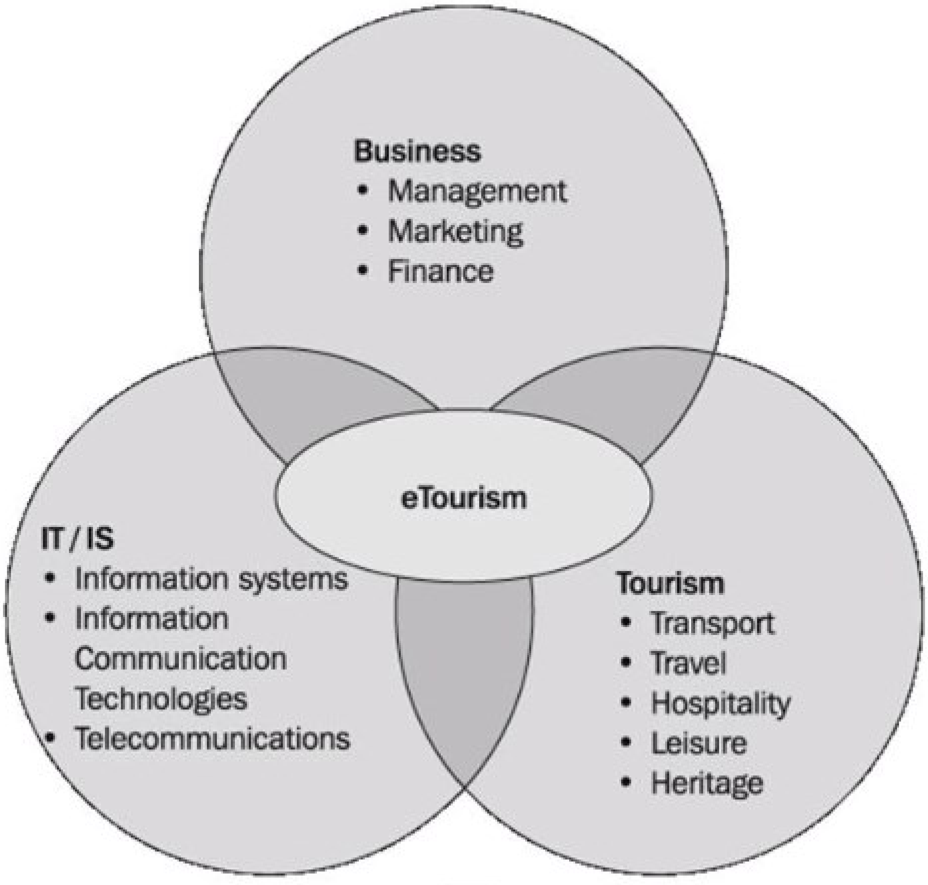
Introduction

Information and Communication Technologies (ICT) are significantly changing the face of many areas of human activity, including tourism. ICTs enable tourists to make online cashless payments, book and purchase products and services easily and efficiently at any time of the day or night, which they can also research in advance for reviews from other customers. Tourism has therefore been developing rapidly in recent years in terms of quantity and quality, and ICT is changing the way services are created and delivered - a trend associated with increasing demands for quality, accuracy, and speed of service delivery, including online and personalised services based on customer preferences, needs and behavioural patterns.

eTourism:

The very close interconnection of tourism with ICT in recent decades has created a new term called eTourism. It refers to the digitisation of processes across the whole tourism sector and infrastructure. **eTourism is developing rapidly and bringing new approaches, practices, and technologies to tourism, which is thus changing significantly in many aspects.** The dynamic development of eTourism is driven by the economically important tourism industry and the rapid technological development that is moving more and more activities to an online form that makes the offer of services and up-to-date information easily accessible in time and at any location. This provides an opportunity for global competition for small tourism businesses as well as for public administrations. The interconnection of different information systems is another driver for the successful development of eTourism. Thus, not only large territorial units such as Europe, but also macroregions (e.g., Visegrad Four) or microregions (e.g., Bohemian Paradise) are conceptually supported, not only in terms of destination presentation. eTourism is also part of the so-called eCommerce, i.e., trade between companies and customers using ICT, while eCommerce also includes the widely used electronic payment system, called ePayment.

The image below illustrates the combination of the different sectors that make up the pillars of eTourism (the previously used acronym IT/IS aka Information Technology and Information Systems is now represented by the more modern acronym ICT):



Source: https://slideplayer.com/slide/5924160/

mTourism

The mass development of the use of mobile ICT (especially mobile devices, mobile applications, and mobile Internet) is leading to the formation of the next stage of eTourism, the so-called mTourism. The spread of these technologies among the public and the associated development of m-tourism brings new opportunities for tourism participants, which are reflected in their activities and, as a result, in their consumption behaviour.
For example, tourism participants in the last decade are assessed by the source (VAŠKO, 2015) as "well informed, without the need for expert advice, more influential, more demanding, more individual, independent and mobile and strongly influenced by the available technologies". With the possibility of high mobility and at the same time constant access to extensive information, today's consumer can make flexible decisions and actions when purchasing tourism products and react to changes when consuming them. As a result, today's tourism participants have less and less need for additional information from tourism product intermediaries. Information resources used for tourism purposes are often filled with information by tourists themselves, a typical example are geosocial networks such as Foursquare, Facebook or TripAdvisor. Information such as satisfaction ratings for products or services are thus obtained by users from other tourists, which often significantly influences consumption or trips of other participants. This user feedback from geosocial networks also has an impact on the quality of tourism products. Evidence for this claim comes from a study by

TripAdvisor (BUSINESS INSIDER UK, 2015), which documents the fact that hotel owners who respond to customer comments are 20% more likely to make additional bookings than those who do not. In Ireland, hotel managers then responded to TripAdvisor customer reviews by correcting deficiencies, resulting in positive feedback and more frequent customer visits.
Participants in mTourism are therefore now passively and actively using mobile ICT, particularly geo-social networks. These are not only a key information source for them, significantly influencing their consumption behaviour, but also an open book of wishes and complaints, which even influences the quality of tourism services. Geosocial networks are therefore an extremely important component of mTourism today.

mTourism is also part of a field called mCommerce, or business-to-consumer commerce using mobile ICT. Besides others, mCommerce also includes the widely used mobile payment system, or mPayment, in the tourism industry.

mPayment – electronic systems of cashless payments via a mobile device (smartphone, tablet, wearables device) of smaller amounts, usually within hundreds of Czech crowns, which are also referred to as micropayments. The following types of micropayments are currently available in the Czech Republic:

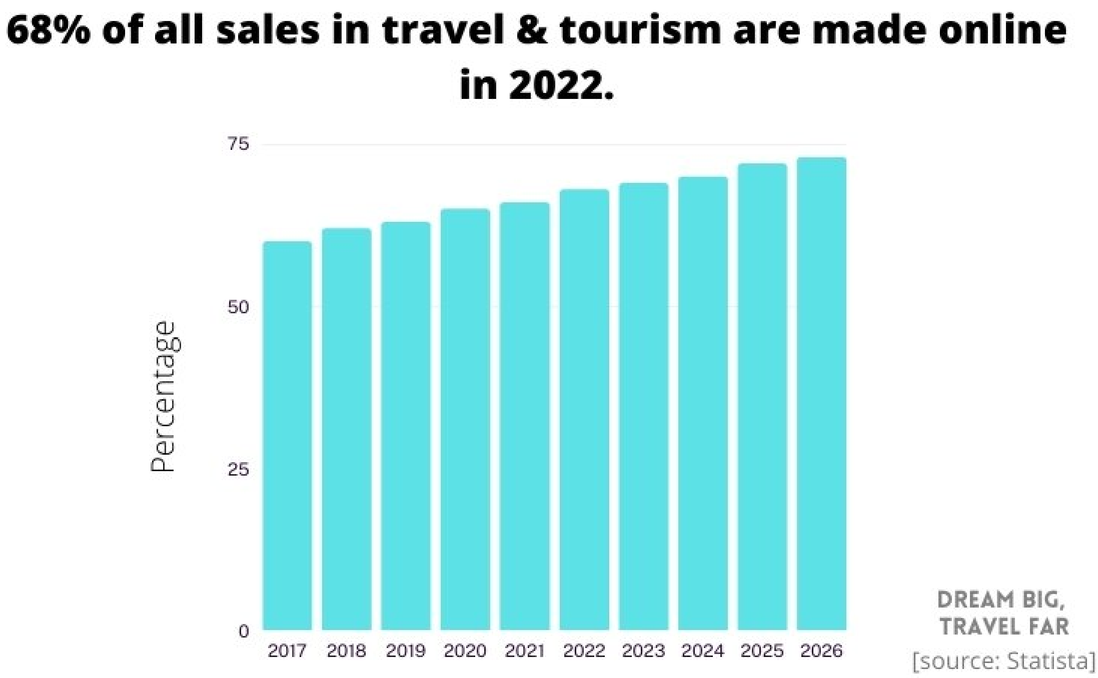
- **Premium SMS** – universal mobile payments, operating with all Czech mobile operators (T-Mobile, Telefónica O2, Vodafone and Nordic Telecom). They are SMS messages with higher tariffs, allowing payment for services or goods for smaller amounts of money, usually in the range of 3 CZK to several hundred crowns. The so-called "Mobile Originated" (abbreviated MO) charges for outgoing SMS messages and "Mobile Terminated" (abbreviated MT) charges for incoming SMS messages. In the Czech Republic, the MO variant is used. The disadvantage of Premium SMS is the very high fees for the merchant to receive payments (about 50% of the amount paid)!
- **Donors Message Service, DMS** – similar to Premium SMS, but only in the amount of 30/60/90/190 CZK and as a payment fee, according to the agreement with the association Forum of Donors (which represents this project in the Czech Republic), mobile operators charge only 1 CZK. However, this type of payment is intended exclusively for non-profit organisations as a way of collecting donations from donors.
- **NFC payments** (Near Field Communication payments) – a modern method of payment with many advantages, using wireless NFC technology of short range (up to about 20 cm). Mobile devices can act as a virtual "Wallet" (the most popular include Google Pay, Apple Pay, etc.) => the user brings the NFC-enabled mobile device close to the payment terminal to complete the transaction. If coupons or discounts are available, they can be automatically deducted from the transaction. This is a very fast and simple payment method, and unlike contactless cards, NFC payments can be turned off on the mobile device.
- **QR code payment** – basically, it is a simplification of payment via mobile banking, i.e., the user scans a QR code from which the payment data is automatically written into a

bank payment order, which the user then just checks and sends to the bank via internet banking.

Innovation of tourism through ICT

Tourism has undergone significant quantitative and qualitative development in recent years, mainly due to the influence of ICT which naturally in the form of eTourism/mTourism integrates into many levels. Sources [1] identify the following as key changes in the tourism industry:

- **Quantitative (long-term continued global exponential growth)**
    - **Searching, creating, and offering new destinations on the web and other media** – ICT enables the promotion, presentation, comparison, and sophisticated search of destinations using many different aspects, approaches, sorting parameters, presentation of outputs, creation of itineraries, etc.;



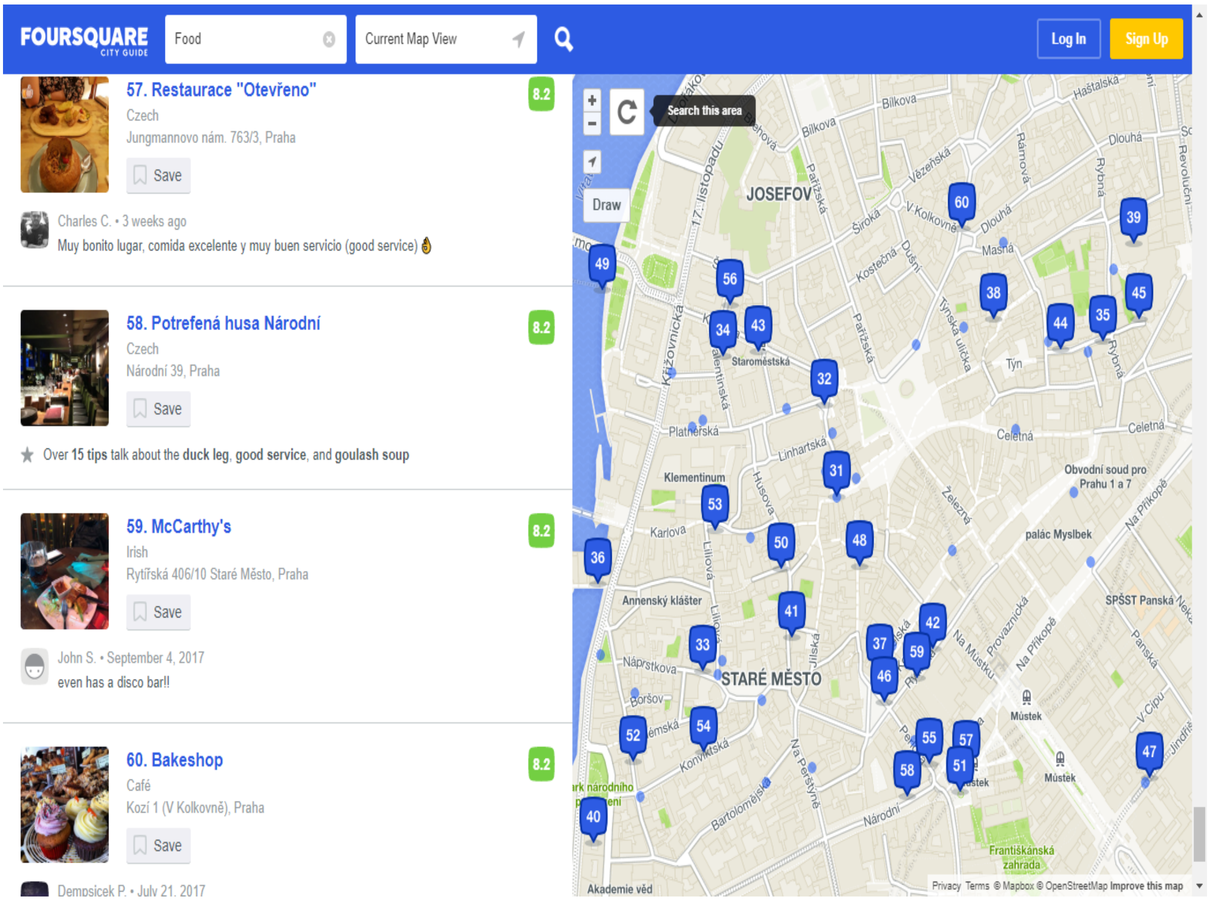
Source: https://www.dreambigtravelfarblog.com/blog/online-travel-booking-statistics

    - **Growing number of domestic and international tourists and visitors** – ICT generates, evaluates and presents statistics; implements automated monitoring; enables modelling of flows and numbers of visitors and development of tourism infrastructure, forecasting of tourism impacts;

- **Engaging a growing number of residents in tourism** – ICT supports the offer for small business (e.g., database of small accommodation facilities, presentation of local services) including the creation of regional packages "recognizable" by the visitor (e.g. thematic trails with the offer of services on the route).

- **Qualitative**
  - **Improving the quality of tourism services** – increasing the quality of tourism services through ICT which enables online availability of many services 24/7 and therefore eliminates delays, improves the speed and reliability of check-in by minimising errors caused by the human factor, ensures the security of services through various types of user authentication, promotes standardisation of services, enables comparison of the quality of tourism services, etc.;
  - **New products offered, including products supporting the trend of sustainable tourism development** – ICT is an opportunity to promote sustainable tourism products (small-scale products, expensive promotion in non-electronic media compared to the promotion of e.g., large resorts), quick availability of information about new products;
  - **New channels of offering tourism products and their combination in communication with potential clients** – in addition to HDTV, ICT offers new channels for offering tourism products - web space, Location Based Services (LBS), on-demand offers via email, offers on data carriers (e.g., multimedia presentation of the destination on DVD/Blu-ray/Flash Drive);

Example of an LBS application, namely the Foursquare geosocial network containing reviews of so-called points of interest (restaurants in the image below):



Source of the image: foursquare.com

- **New ways to offer products** – various types of packages, discounted product offers to "regular" clients, use of new technologies, many directly involving ICT – (large screen) projections on transport terminals, use of augmented and virtual reality, multimedia, interactive presentations, 3D cinemas, webcams, etc.

Example of an application (available at www.nezmizelo.cz) using augmented reality to visualise the no longer existing demolished monuments in the city of Pardubice.

Source: www.nezmizelo.cz

- **Temporal**
    - **Efforts to reduce seasonality and higher use of tourism infrastructure capacity, extending the season** (snowmaking, complementary offer of activities in the destination - attractions) - using ICT to offer off-season products, creating off-season attractions (e.g. 3D cinema for "visits" to nature in the off-season);
    - **Congestion of traffic roads and premises in the season** – ICT supports monitoring, traffic management, modelling of passenger flows (e.g. models for airport terminals) and thus prevents/reduces congestion;
    - **Immediate handling of requests from clients and tourism entities** thanks to ICT at check-in and access to services via e-business (GDS, CRS, web presentations of tourism services, attractions, regions, etc.);
    - **Aiming for long-term sustainable tourism development** – ICT has the potential to monitor and manage visitor flows, monitor and evaluate changes to destinations within the destination life cycle, evaluate and model tourism impacts, create a basis for the use of destination life cycle concepts, carrying capacity, limits of acceptable change.
- **Sustainability of tourism**
    - **Sustainability of tourism becomes a priority in protected areas and areas of very high cultural value (UNESCO monuments and others.)** – ICT supports communication with the participants (visitors, locals, investors, etc.), it can mediate the "virtual accessibility" of cultural heritage (e.g. rare artefacts otherwise inaccessible to the visitor – e-culture);
    - **Tourism development and its impacts are monitored, evaluated, managed** – ICT supports the evaluation of data on the course of tourism as part of marketing

research and monitoring in the context of sustainable tourism development, an important tool is GIS for monitoring the development of the territory (flora, fauna, environmental components), registration of natural and cultural monuments, support for the derivation of the load-bearing capacity of the territory, delineation of various zones in relation to tourism, documentation for EIA, optimisation of the management of hiking trails and routes, optimisation of the location and layout of tourism infrastructure.

- **Knowledge-based decision-making on tourism sustainability** – ICT offers expert and knowledge systems, availability of studies, best practice guides, methodologies, linking of acquired knowledge e.g. in RIMS, etc.

- **Manipulative (marketing)**
  - **Advanced methods of acquiring new clients** (new types of packages with flexible length, content, range of selected and paid services, "suitable name" - e.g. management stays, linking the offer to customer loyalty programmes, feedback from previous clients, etc.) - ICT is a medium of promotion, linking, flexible choice (e.g. if I choose this package, how much will it cost...);
  - **Acquiring regular/frequent customers** (frequent flyer programmes and other loyalty programmes, ensuring client satisfaction and positive feedback, good image, public relations, etc.) - ICT is the medium of "loyalty" communication (introducing loyalty products, checking that participation in services is included in the client's programme score, selecting programme benefits, etc.), enables sophisticated marketing analysis of loyalty programme data (e.g. changes in the loyalty programme on client behaviour and interest), creates a comprehensive image of the tourism entity;
  - **"managing" opinions of tourism participants** (groups of interest - especially potential clients, tourism service mediators, tourism service providers) - ICT is an important part of all marketing tools, it supports the implementation of marketing research, modelling in marketing (e.g. forecasting demand for tourism services);
  - **Creating demands for destinations and products** – ICT promotes tourism products, types and forms of tourism and destinations in many forms (visual and auditory impact, combination of stimuli, their different processing, etc.), in different contexts (complex offer, targeted search, logical classification of the offer, offer in thematic context - e.g. castles and chateaux, etc.), using different technologies (different media, different interactivity, different resolutions of the offer - e.g. according to the size of the display for presentation, etc.), in different locations including online in the field (LBS), in different structuring and quality of the offer according to the subjects producing the offer (national tourism headquarters, local tourism authorities, enthusiastic individuals, etc.).

- **Emotional (perceptual, sensational, experiential)**
  - **Seeking extreme experiences and the related building of tourism infrastructure and human resources for them** – ICT can provide "safe or

unreachable" extreme experiences (virtual reality travel to space, remote destinations, mountain peaks, etc.), educating and training guides;

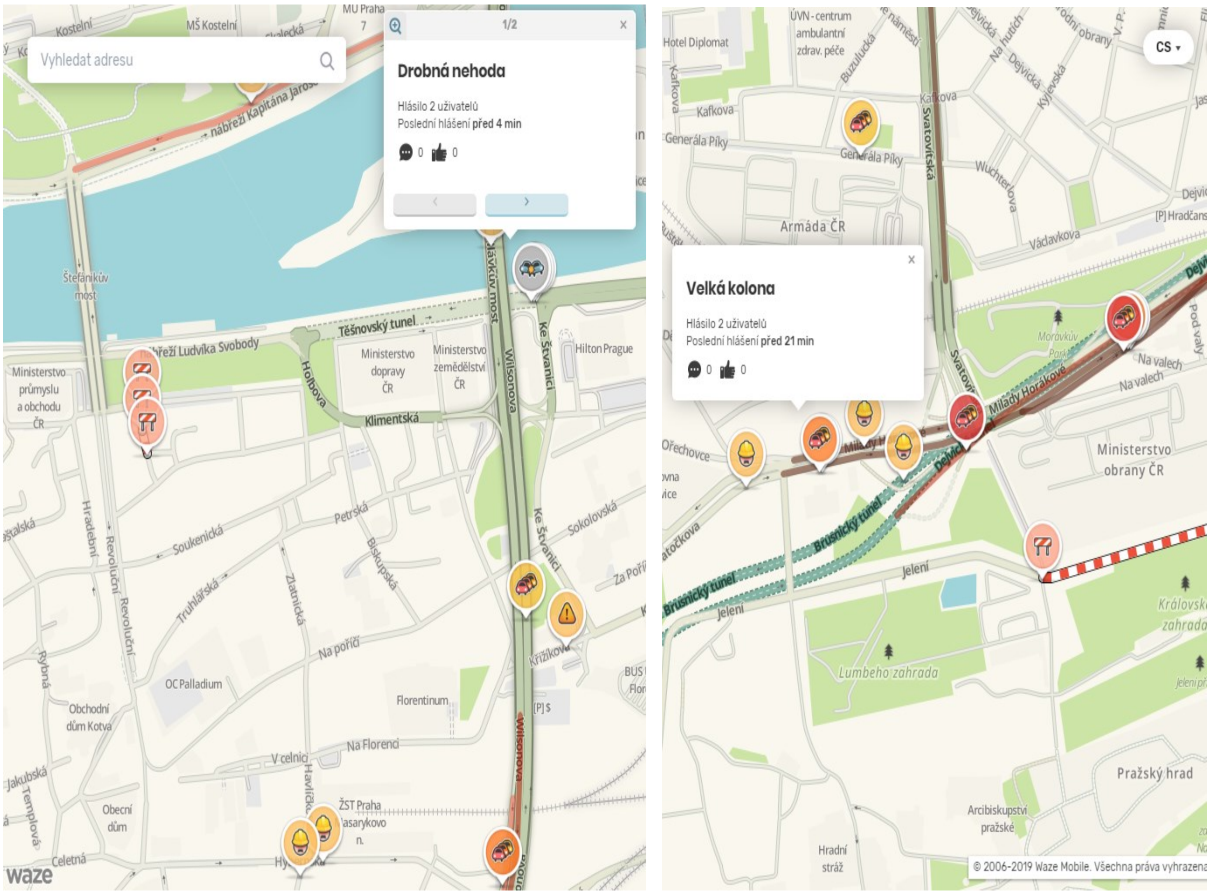
- **Seeking „intimate“ experiences** (quiet sitting without holiday stresses in a flowering mountain meadow, sunrise in a mountain valley, etc.), visitors gradually find a way of getting to know and meeting the cultural and natural landscape - ICT can bring quality motivational photos, quality video and audio information, discussions of tourists about the focus of their holiday, the benefits of an experiential (and educational) holiday, etc.

- **Cognitive (human thinking)**
  - **The thinking (and preferences, behaviour, priorities, way of experiencing) of visitors is changing**, the result is, among others, a gradual change in the traditional 4S formula for seaside stays to exploration and closer contact with the local population, a trend towards more emphasis on the interpretation of landscape and cultural heritage, the growing attractiveness of ecotourism - ICT promotes "communication" with local natural and cultural heritage, exploration and interpretation of the landscape, knowledge of the risks of a 4S holiday and all this to change this traditional formula;
  - **There is a growing demand for individual (or small group) exploration of rural and natural areas, including interpretation of different aspects of the destination** – ICT supports the search for small monuments, attractions and local small tourism services in the region (crosses, local museums, eco-farms, etc.), creates virtual educational and interpretive trails, can provide information and knowledge targeted to different client segments;
  - **Beginning of the use of cognitive and mental maps in promotion** (the way of offering is more close to the typical features of human perception of space) - ICT makes it possible to create mental maps (e.g. a mental map of a region working with a combination of symbols and geographical data), to use them for linking to other information (a mental map as an interactive map);
  - **Visitors' responsibility for the development of the visited destination is increasing** (the visitor is a "guest" of the visited community, destination, nature and landscape) - ICT prepares visitors to behave responsibly by providing information on the visit, codes of ethics and codes of behaviour, and the potential impacts of their presence.
- **Information-communication**
  - **Growing geographical** (LBS, linking to the web and other media), temporal, expertise and preferential (adapting to the end client through communication channels, content, service, design, etc.) availability of information about tourism and tourism services;

Source: https://www.openpr.com/news/2013699/what-are-the-chances-of-location-based-services-market-to-grow

- **Growing flexibility of IS in many aspects** – design, information content and its interconnectivity, applied technologies, ways of logical classification of information, ways of information search, etc.;
- **More and more information online** – webcams (destinations, means of transport), current weather, snow cover, current occupancy of accommodation and means of transport, current traffic connections and traffic situation, arrivals of means of transport, etc;
- **Use of sophisticated IS** – increasing complexity of tourism service offerings, higher proportion of graphical information and graphical interfaces (GIS, interactive maps, panoramic maps, traffic maps, multimedia information, etc.), multiple methods of accessing information, structured information, multiple layers of information, etc;

Image from the Waze geosocial network designed to share the current traffic situation on the roads - thanks to the information of many users (in the Czech Republic, about 700,000 drivers use this network), the maps provide very up-to-date information about traffic jams, traffic accidents, road patrols and other restrictions:



Source of the image: waze.com

Mobile ICT

= information and communication technologies providing users the ability to be mobile, i.e., to move around in the field, which is crucial for tourism participants. This is possible thanks to wireless and mobile data technologies, mobile devices (smartphone, phablet, tablet, laptop) and, of course, mobile applications.

Mobile devices

Mobile ICT refers to those devices that provide mobility capability to the user, are not tied to a specific location - they operate without a mains power source, network data transmission is managed wirelessly - preferably over mobile data networks, etc. Mobile devices allow tourism participants to obtain instantly - thanks to their high mobility even anywhere in the field - up-to-date information and suggestions on tourist destinations during the travel planning process.

Mobile devices suitable for tourism purposes:

- **Smartphone**

- **Tablet**
- **Phablet** – a hybrid between smartphone and tablet, 5-7" in size (large enough touchscreen for web and photos but compact enough to fit in a pants pocket).
- **Laptop** – is a general term for the first models of portable computers, usually larger in size and weight than those commonly used today.
- **Notebook**
- **Netbook –** a portable computer of very small size, usually with lower power, designed for easy and fast connection to the Internet and web or email browsing. Not very suitable for more demanding applications.
- **Ultrabook** – a designation from Intel (2011), equipped with its own CPUs. Higher mobility due to minimum requirements for thinness and weight of the computer/speed of waking up from sleep mode/battery life/CPU architecture/disk transfer speed/availability of ports or software equipment. Intel released the specification in 2011 as Huron River, in 2012 as Chief River, and in 2013 as Shark Bay. The very low weight (less than 1.5-2 kg) is countered by the absence of some components (e.g., DVD drive).
- **Convertible notebook** (or **Ultrabook**) or **2-in-1** – a hybrid device with a classic notebook design that can be transformed, for example, by means of a flip-over, in-frame rotatable, detachable, or plug-in display into a touchscreen-controlled tablet. The 2-in-1 device thus removes the boundary between laptop and tablet.

Examples of a convertible laptop with a flip-up display and multiple usability positions:

- **2-in-1 with detachable keyboard** is suitable for users who need easy portability of the device (e.g., in the field when navigating in tourism), but still want to be able to work and type on the computer more easily when the keyboard is connected (e.g., when they arrive at the hotel from the field).

  Advantages:
  - When the keyboard is removed, it can be used as a small and light tablet.
  - Quickly transformable into a tablet.
  - In some cases, the keyboard also has its own battery or storage.

  Disadvantages:
  - When the keyboard is removed, it must be carried separately.
  - The part of the device with the display can be significantly heavier as it contains most of the 2-in-1 hardware.

Source of the image:

https://www.gizmodo.com.au/2015/06/the-new-tech-you-need-to-understand-before-buying-a-new-2-in-1-laptop/

- **The 2-in-1 with 360-degree swivel display** is suitable for users who require working with a full-size laptop, but with the ability to transform for greater comfort when watching videos, for example. For the needs of travellers, this one is therefore less suitable.

  Advantages:
  - Robust, easy to transform.
  - Holds the display in every position without additional support (necessary with a tablet).
  - With up to 360 degrees of rotation, the display can be adjusted to any angle (usually a laptop can be rotated between 120-160 degrees).

  Disadvantages:
  - When the 2-in-1 keyboard is turned into a tablet shape, its keyboard lies on its underside, which is placed on a table, etc.
  - Significantly lower mobility for the user than the previous type.

Source of the image:

http://images.bestbuy.com/BestBuy_US/en_US/images/abn/2012/com/pcon/lenovo/hero_image4.jpg

Mobile device classifications:

- **Single-purpose mobile devices** – provide a use for one specific need, such as a GPS or modem for wireless and mobile data technologies.

Source of the image:

https://www.campingmultistore.de/tomtom-go-camper-navigationssystem-00636926100571/

- **Multi-purpose mobile devices** – integrate multiple single-purpose device technologies, e.g., smartphone, tablet, laptop.

Source of the image:

https://www.itwire.com/mobility/phablets-poised-for-local-growth-too.html

Augmented reality in e-Tourism

Jiří Kysela and Pavla Štorková

Augmented reality is a relatively new type of user interface for mobile applications that enables a mobile device's camera to capture a real image of the user's surroundings, place digital objects in it, and then display everything together on the mobile device's display (a smartphone or tablet is suitable). Compared to virtual reality, the advantage of augmented reality is that it does not isolate the user from the surrounding world with any headset (special virtual reality glasses that completely obscure the view of the surroundings) and, on the contrary, it leaves him the perception of the surrounding real world, which is enriched by a digital layer displaying objects generated by the application. The digital objects can be represented by static images but also by video footage.

Among the public, the most successful application in recent years has been Pokemon Go, which is based on the use of augmented reality. This type of interface allows tourism participants to interact with their surroundings by displaying visual content in an innovative way. Augmented reality is ideally applied in outdoor use when the user of such an application is mobile.

Another example would be the visualization of now vanished historical buildings in their original form and in the exact location where they once stood. This method of use in tourism is illustrated by the images in the text below, where the no longer existing historical buildings in the cities of Pardubice and Litomyšl are visualized through an augmented reality application.

Technology today also offers teachers many opportunities to engage students and offer them innovative ways of presenting the material they are studying. The development of technology is constantly moving forward and so teachers can also work with new technological tools. With today's capabilities of computers, tablets, and mobile phones, not only is it possible to display the material, but with the help of augmented reality and the appropriate equipment, students can directly participate in interesting experiments. Any student who owns a mobile device (tablet, smartphone) can work with multimedia content enriched with augmented reality. Augmented reality has great potential as a means of highlighting interesting features or bringing history to life.

New media

Augmented reality is most often presented through new media, which include mobile applications that can use image sensors (camera) and motion sensors (accelerometer, gyroscope, electronic compass, geolocation sensors) available on mobile devices to provide users with digital graphic content placed exactly in the desired location of the real image. The new media can be defined by the following rules:

    - New media is **based on an electronic platform.**
    - They use **the calculative power** of computers.
    - They are **interactive**.
    - They encourage **communication and feedback.**
- Typical **characteristics of new media are:**
    - **Multimediality** – they integrate text, image, sound, video, etc. through the computer.
    - **Interactivity** – allows the user to intervene in the processes and react to the information obtained.
    - **Virtuality** – they allow users to create their own or shared abstract "cyberspace".
    - **Communication** – very easy communication between individuals or groups.
    - **Globality** – see "global village", overcoming geographical barriers.
    - **Internationalization, distribution, diversification, and mobility** – independence of place, time, and other conditions.
    - **Collaboration** – easy collaboration in virtual teams or communities.
- **Typical new media include geosocial networks** (e.g., TripAdvisor, Facebook, Twitter, Foursquare, Instagram, etc.).

Smart devices and smart tourism:

Smart mobile devices on which users can run augmented reality applications are the cornerstone of augmented reality. These are devices that are equipped with a mobile operating system (Android, iOS, etc.) and that can communicate via a computer network and interact with other smart devices. The most well-known representatives of smart devices that can be used for augmented reality include, of course, the smartphone, the tablet and, in recent years, the increasingly widespread socalled wearable electronics (small electronic devices designed to be worn on the body) which, when used for augmented reality, takes the form of a so-called headset with glasses, speakers and microphone (e.g., Apple Vision Pro).

All these technologies are the basic cornerstones of so-called smart tourism, which depends on key mobile ICT technologies such as smart devices with mobile operating systems and mobile applications, and of course, wireless, ideally mobile Internet is now essential for the ability of such devices to communicate. For augmented reality to work, it is necessary that the operating system of the mobile devices used supports it - specifically Android using ARCore technology, and iOS using ARKit technology.

Augmented reality in tourism

Unlike traditional tourist information channels, such as information centres, an augmented reality app can offer a permanently accessible online source of information, so visitors can educate themselves whenever they want. The user does not need to use information centres, search for information on the Internet or orient themselves in printed maps and brochures. They gain time flexibility and mobility and learn only the information that interests them. On a mobile device, they can view the entire route on a map, which may show points of interest with important historical events. For example, signs with QR codes can be placed at the marked points of the tourist route, or the app can offer the visitor various 3D visualizations (historical buildings, places that no longer exist today that can be compared with reality, etc.) in the form of augmented reality via GPS coordinates. Tourism and education will gain a new dimension in exploring new places and discovering local history, offering much more information in a simple form hidden in a mobile device. Therefore, mobile devices with the support of augmented reality technology offer much more interactive and interesting information in one place in real space-time compared to traditional information sources.

Mobile web app with augmented reality

The objective of the project "Activation of participants through social networks and gamification in the areas of undertourism" implemented with funding from the "Iceland Liechtenstein Norway grants" in the years 2021-2023 at the University of Pardubice, was to develop a mobile web application with augmented reality.

- The project focuses on the promising IT sector (**augmented reality, social networks, and gamification**) => through which it aims **to activate interest in local tourist points of interest in the Czech Republic.**
- The aim of the project is, among other things, to create and operate a mobile web application with multimedia content for the needs of tourism and **education of students and the public.**
- The application will offer the user (tourists, students) local **information about the visited points of interest virtualized through augmented reality and 3D models of monuments.**
- The application is also used by the city Tourist Information Centres (TIC Litomyšl, TIC Pardubice) and the regional Destination Company East Bohemia. The use of the application by these institutions is for the purpose of supporting gamification - both TICs have access through another auxiliary web application to a database with users of the main mobile web application and information about which sights tourists have already visited

or have checked in at their location. Based on this, the TICs can check whether tourists who visit the TIC and claim a gamification reward have already fulfilled the condition of visiting all the monuments in each city and can therefore provide a free reward. Specifically, souvenirs with the theme of the given destination are available as rewards - e.g., in Pardubice a porcelain mug or a bell with the theme of the monuments of Pardubice.

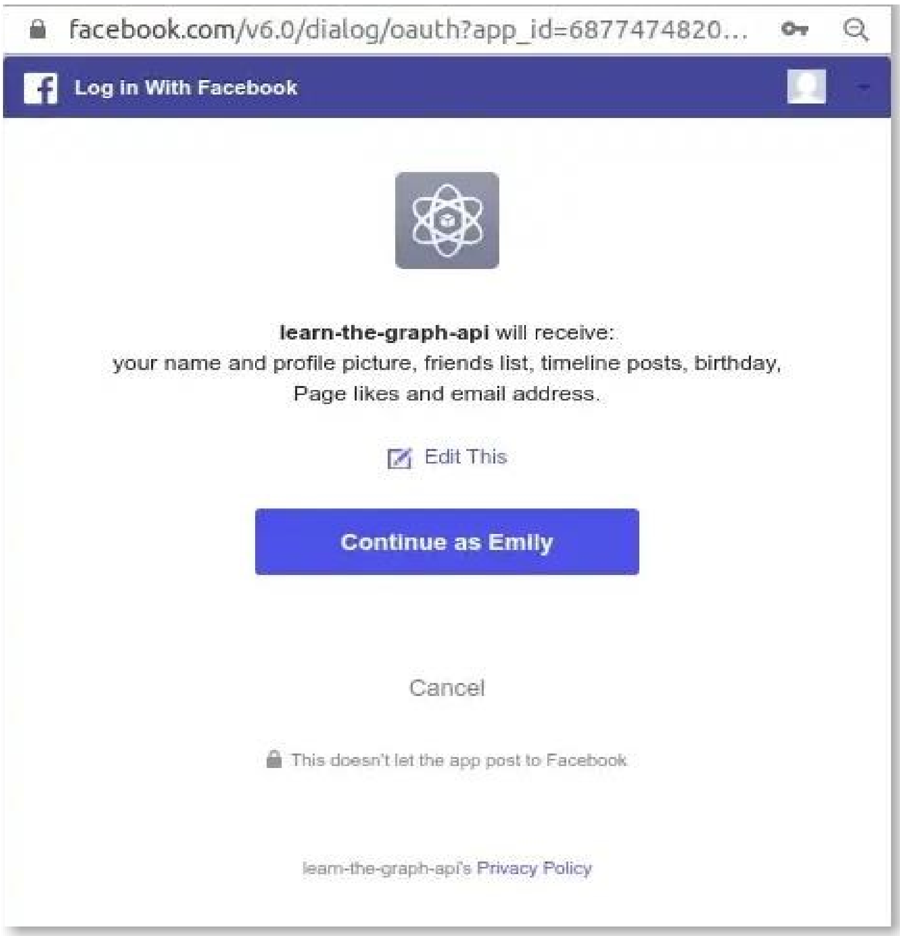
- By promoting the app, they get a means of tracking **the flow of tourists/data about them:**

  date and time of visits, country and, thanks to Facebook gamification, gender, age, city of origin, etc.

  A list of all data that can possibly be collected via the Facebook API about users logged in via Facebook login:

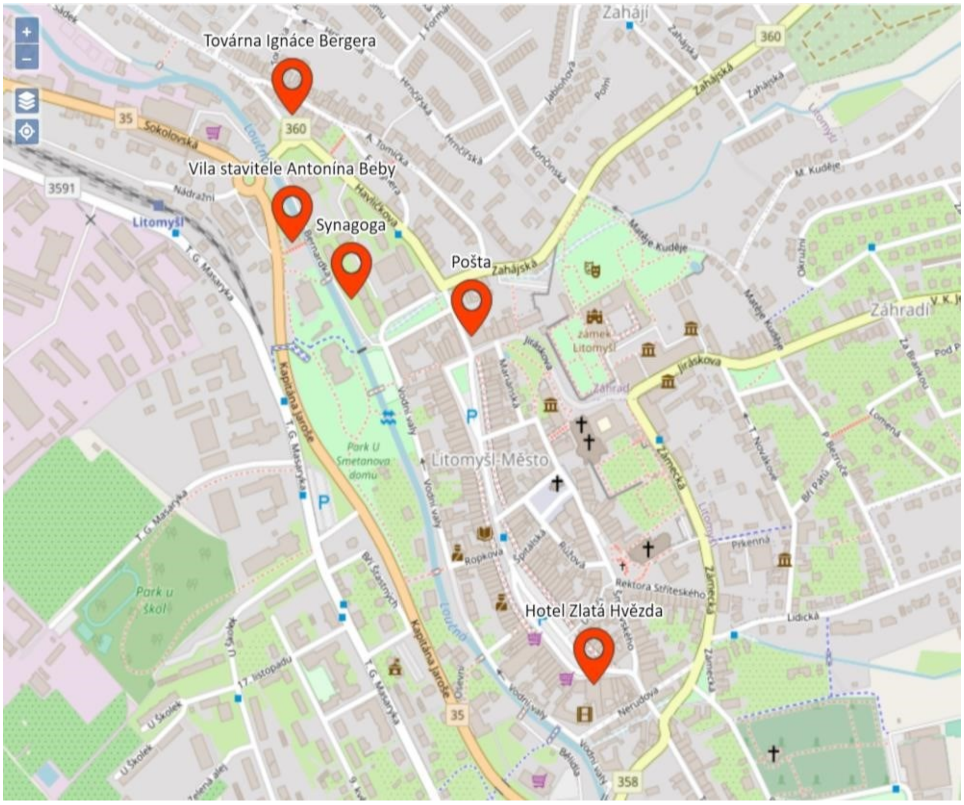
  - user_age_range
  - user_birthday
  - user_friends
  - user_gender
  - user_hometown
  - user_likes
  - user_link
  - user_location
  - user_messenger_contact
  - user_photos
  - user_posts
  - user_videos

A screenshot of the Facebook login form used within the custom web application:



Source: https://levelup.gitconnected.com/get-acquainted-with-facebook-login-and-the-graph-apid1e951e0a1da

- The project covers the regional city of Pardubice and the UNESCO town of Litomyšl.
- The app is **free and available on all platforms** (Android, etc.) that support augmented reality technology. Unfortunately, the iOS operating system's support for augmented reality in web applications (the so-called WebXR standard) is still limited in August 2023, which does not yet allow augmented reality to work.
- The application is available directly on the web in the browser **without the need to download and install** from the app store (App Store, Google Play, etc.).



- In addition to informative texts, the mobile device also displays **multimedia content specifically designed for the area.**

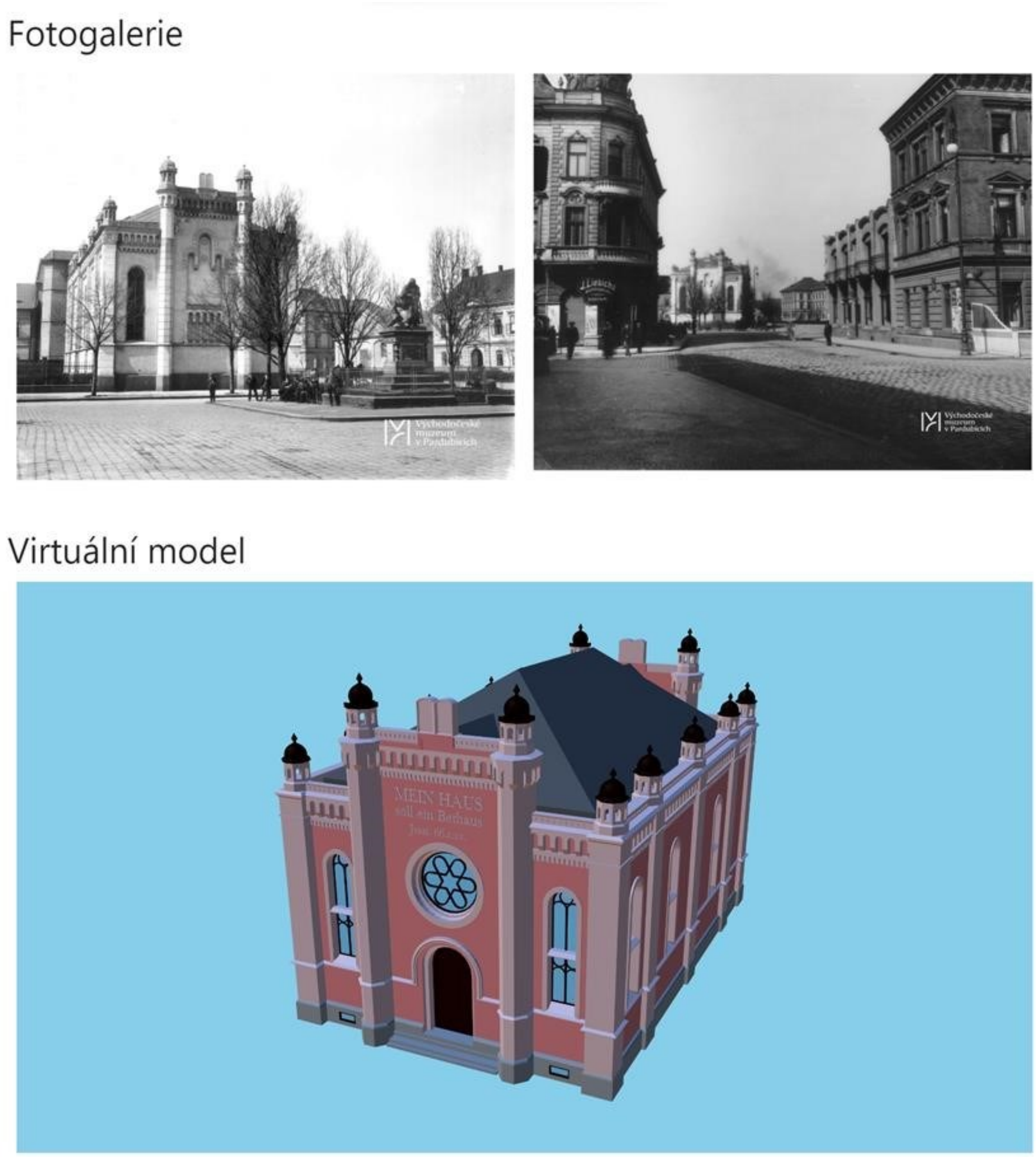
- Thanks to augmented reality, the user can see digital **3D graphic models of historically significant, now vanished buildings on a smartphone/tablet (reconstructed by graphics thanks to historical building plans from museums/archives and historical photographs).**

- The tourist can **compare these models in the locality of their actual historical occurrence with the reality in front of him in the context of the surroundings**, i.e., the current construction.

- The reconstructed monuments include, for example, **extinct synagogues, the town hall, the castle gate, and other important historical buildings.**

- A mobile web application of this type is not yet present on the Czech market.

Screenshots capturing the details of the monuments in the mobile web app created during the project:



Source: www.nezmizelo.cz

The screenshots below show the augmented reality mobile web application created during the project:


DŮM SLUŽEB
TOP SHOP


HOTEL
VESELKA


MĚSTSKÉ MUSEUM
Městská spořitelna

Source: www.nezmizelo.cz

Sources:

- www.nezmizelo.cz
- KYSELA, J., ŠTORKOVÁ, P.: *Using augmented reality as a medium for teaching history and tourism*. In: INTE 2014 Proceedings Book. Sakarya: Sakarya University, 2014. ISSN 2146-7358.

- KYSELA, J.: *Na cestách po bodech zájmu, aneb geosociální sítě v roli nejlepšího turistického průvodce?* [online]. Available at: http://www.internetprovsechny.cz/na-cestach-po-bodechzajmu-aneb-geosocialni-site-v-roli-nejlepsiho-turistickeho-pruvodce/

Gamification

Jan Panuš

Introduction

Gamification, which is referred to as the marketing word of the day, can - and does - mean different things to different people. Some see it as creating games that are explicitly designed to promote products or services. Others picture it as creating 3D virtual worlds that lead to behavioural change, or a method of training users in complex systems.

They're all right. Gamification connects all the disparate threads that have been developed in games for non-game contexts. Thus, we bring together serious games, and games for the change into a coherent worldview that is based on the latest research on the behaviour and success of social games. For our purposes, we will define it as follows:

Use game thinking and game mechanics to engage users and solve problems.

This framework for understanding gamification is powerful and flexible: it can easily be applied to any problem that needs a solution.

Players Motivation

The essential element of gamification is the player. In any system, how the players are motivated determines the results. Understanding player motivation is therefore crucial to building a successful gamified system.

We already know that games are generally good motivators. By focusing on three main components - pleasure, rewards and time - games have become one of the most powerful forces in all of humanity. What is unique is that games can predictably get people to take actions that are not always in their best interests, without using force.

Powerful human motivators

From Greek mythology to daytime soap operas, it's clear that sex - or the desire for it - will get a person to do almost anything. Paris' kidnapping of the beautiful Helen of Troy led King Menelaus to start the Trojan War. So, like games, sex has the unusual ability to make people do things that are not in their best interests. But like the Trojan War - and unlike games - it is not a predictable motivator.

Similarly, violence can produce unprecedented motivational results. Putting a gun to someone's head is likely to get them to do whatever task you want them to do. However, he is not likely to enjoy a second of it - and he certainly won't be back for more. The mistake about the power of punishment to achieve great results is a strong and wrong belief.

Games, however, hit the mark. They combine the desire for sex with the predictability of pressure - only without using force, and when they are successful, they are driven solely by pleasure. This pattern is also why the games have a dark side: People addicted to slot machines can look like they haven't seen the sun in months, while World of Warcraft players are sometimes accused of neglecting their real-life responsibilities because of the game. But there is a bright side to games - that they change people's health, their way of learning and their way of life for the better.

Flow

Something called flow is fundamental to the success of games. Our understanding of flow is based on the research of Mihaly Csikszentmihalyi, an American professor of psychology of Hungarian origin, who became famous for his work in the study of happiness and creativity. Achieving flow - or being "in the zone" - refers to a state where the player is between anxiety and boredom and meets their own motivational level in this experience.

When a jazz player plays music or a runner trains, they exist in a state of suspended animation. They are calm and focused. The writer, in the midst of the flow of the narrative, forgets for a moment the world around her/him. It is safe to assume that almost everyone has had the experience of losing track of time and space while playing a game, cooking, exercising, cleaning, or making a phone call.

Game developers, meanwhile, are obsessed with creating this state for their players. They are constantly looking for ways to get the player to merge with the game. It's a constant effort to bring someone into their system only to leave them seamlessly in a highly valued state of flow - but how?

Through careful interplay between system and player, and a relentless belief in testing those interactions to find that point between anxiety and boredom. And with a series of other psychological phenomena, such as reinforcement.

Reinforcement

If a mammal, such as a lab rat, receives food pellets at a fixed interval every hour - during the 59 minutes between pellets - the animal will always leave and start doing something else in the cage. Only at the 60th minute does it return for the pellet it was given.

This structure is similar in form to many industrial-era workplaces. The worker is paid every two weeks. What happens in between is entirely consistent with this end result. In other words, the worker will do exactly what is required of her/him to ensure that she/he receives a bi-weekly paycheck in the days in between. Nothing more, nothing less. This is called fixed-interval reinforcement. Not surprisingly, fixed-interval reinforcement schedules tend to yield low levels of engagement.

At the other end of the spectrum is strengthening with a variable ratio and variable schedule. In this model, the lab rat does not know how big the reward will be or when it will come, but it knows at some point that it will come. Therefore, the rat will endlessly press the dosage pedal in the cage until it receives the reward. This is exactly the model used in slot machines and almost all other archetypal models of gambling. Another name for this behaviour modifier is operant conditioning, and it is undeniably addictive to mammals.

While operant conditioning experiences can be dangerous and disabling for many, their reasonable use within a broader game-like experience is a powerful force for driving player behaviour.

Unlike caged rats, however, most people are not forced to play - in fact, such a design can be a big discouraging factor for engagement. So why do we play at all?

Why do people play

A good working theory of why people are motivated to play games argues that there are four basic reasons, which can be seen together or separately as individual motivators:

1. To manage
2. To relieve stress
3. To have fun
4. To socialise

Nicole Lazzaro, an expert on gamers' experiences and emotions in games, described four different kinds of fun in her 2004 article "Why We Play Games".

<u>Hard Fun:</u> in which a player tries to win in some form of competition.

<u>Easy fun:</u> the player focuses on exploring the system.

<u>Altered state:</u> when the game changes the way a player feels.

<u>Social fun:</u> during this, the player engages in a game with other players.

In 1964, the famous social science bestseller "Games People Play" by Dr. Eric Berne was published, revealing games organically cultivated through social interaction. The book, which focused heavily on the social engagement of "housewives" (certainly an outdated but fashionable construct for its time), managed to recognize some interesting insights about social gaming. In one of the book's most compelling examples, Berne talks about women's playful engagement, in which they go around in circles and talk about what their husbands are doing to annoy them.

"He leaves socks everywhere," one of them whines. "He hates my cooking," says another. "He forgets my birthday," complains a third.

But what would happen if someone in this circle decided not to play the game? According to Bern's research, if the fourth housewife says, "My husband is actually a good guy," there is a consequence for her that is decisive and punishing. At the next party, the fourth housewife will not be invited back - plain and simple. The reward in this game is to win a subsequent invitation.

An interesting point that Berne leaves us with is an emerging understanding of the social games that exist organically in layers of our society. Even the language suggests that we have been aware of these "hidden" games for some time: one is called a "player" if one can get a lot of dates in the "dating game." A government official "runs" for office in a "contest". Someone who is socially unsuccessful is labelled a "loser."

Types of players

The more you know about who is playing your game - both current and potential players - the easier it is to design an experience that will drive their behaviour in the desired way.
One rubric that can help you understand your players is to use the work Richard Bartle has done in understanding player types. In his seminal work based on his study of MMOG players, Bartle identified 4 types of players. Since then, the number has expanded from four to between eight and sixteen, but these four remain arguably the most stable and therefore the most interesting for our purposes:

Explorers

The explorer simply likes to go out into the world to bring things back to his community and declare, "I discovered this thing!". One example of a game suitable for the explorer type of player was the Super Mario Brothers game on the Nintendo Entertainment System. The player had to play 100 or more games to find every level hidden behind every pipe and block and bring it back to their peers for praise.

Achievers

People who like to achieve are an integral part of any competitive game. They are the driving force behind many projects, services and brands. The problem with designing exclusively for this type of player is that it is difficult to develop a system where everyone can win and succeed. And for players who are successful, losing the game is likely to lead to a loss of interest in playing it.

Additionally, a common distortion that your designers have observed when working with clients to gamify their experiences is that most system, site, and product designers are high-achievers. From this, you naturally infer that most gamified people have similar tendencies. Turns out that's not true at all. Most are socializers.

Socializers

This type of gamer consists of people who play games for social interaction. Games aimed at socializers make up some of the most enduring games in history - dominoes, bridge, mahjong, poker - and what they all have in common is that each is an extremely social experience. But to be clear, it's not that socializers don't care about the game or about winning - they do. But the game is the backdrop for meaningful long-term social interactions. It is a context and a catalyst, not an objective in itself.

Killers

Killers, also known as griefers, make up the smallest population of all player types. However, they are an important group to understand. Like achievers, they desire victory - but unlike achievers, victory is not enough for them. They must win and someone else must lose. Moreover, others have to see it or it wasn't really a win.

Bartle did not intend for these four player types to serve as a personality inventory - they were not developed with that intention. But with decades of gaming thinking behind them, it's easy to see how they can be useful when considering players of gamified systems. And by arranging player types on an axis, we can see how they move from play to interaction and from people to environment.

Note: People are not exclusively one or the other of the four types of players. In fact, most people have some proportion of each. Most likely, a person's dominant type of gambler changes over the course of a lifetime - and even varies from game to game. For a game designer, however, this is a compelling way to see how people are motivated to play and interact in a game system.

If you take the Bartle test to find out your player type, you will find that, as we mentioned, they are mutually inclusive. In other words, a player can be 100% explorer, 100% achiever, 100% socializer, and 100% killer, and all at the same time.

Most people aren't like that. For an average person, the classification might look something like the following:

80 % socializer

50 % explorer

40 % achiever

20 % killer

However, if these ratings were mutually exclusive and a player could only be one or the other, we would learn that the vast majority of people are social: up to 75% of us. In the context of the skyrocketing success of games like Farmville or poker, this statistic shouldn't be particularly surprising. Explorers and successful gamers each make up roughly 10% of the population, and killers 5%.

Combining many of the insights we have developed in this chapter, the foundations of the social games movement suddenly become more obvious.

Social games

Video games, which have influenced much of our gaming thinking in recent years, are in fact the exception, not the rule. Only hits like Farmville and World of Warcraft (WoW) have begun to introduce a social element into gaming thinking. This is despite the fact that most of our games have historically been social. In fact, almost all of our games throughout history - with the obvious exception of Solitaire - required other people to play.

Video games were created as the first true single-player games. So for most video game developers, their view of the importance of socialisation in games is distorted. The last 35 years have been spent designing single-player games. So the success of a simple social game like Farmville seems like an amazing breakthrough, even though it's actually the status quo.

The truth is that game developers and technical designers are unlikely to be in the norm when it comes to the players they design a game for. Just like you, they tend to be a more success-oriented race than an average person. It's a big bias that needs to be overcome. More than likely when they think about game design, they accumulate points, strive for status, and even kill to win. But they are not average people.

Average people are trying to socialise, not win. Winning is not what drives society. If designers start out thinking the game is about success, at some point they realise they're wrong. The average WoW player is as dependent on the guild (team) they fight their battles with as they are on the battles themselves. Excusing himself from a family dinner so he doesn't have to let his team down because of a raid at 7pm is a much greater motivation than the raid itself. The raid itself could wait until after dinner. Most of the company plays more for the comradeship and community of the game than for the victory.

Game Mechanisms:

Design for engagement

Game design is a relatively new, non-accredited discipline with roots in psychology and systems thinking. We use many aspects of game design to create the game experience while focusing on the core elements that will bring the most impact to our players. For example, in gamification, we typically ignore narrative structure because we are creating "non-game" experiences. This means that the arc of your gamified system is based on the player's story and the brand - as they already exist.

Fortunately, you don't have to and shouldn't want to become a full-fledged game designer. While many reference works (such as Jesse Schell's excellent The Art of Game Design: A Book of

Lenses) will help you deepen your understanding of how to make games, we've filtered the key elements of the discipline here and focused on the most important aspects. Our perspective on game design is narrow but highly optimised for gamification.

MDA Framework

One of the most commonly used game design frameworks is referred to as MDA - which stands for:

Mechanics

Dynamics

Aesthetics

The MDA framework represents a post-mortem analysis of the elements of the game. It helps us to use systems thinking to describe the interactions of these game elements and apply them beyond games.

The mechanics form the working parts of the game. Essentially, they allow the designer ultimate control over the levers of the game - they give the designer the ability to control the actions of the players. The dynamics, meanwhile, represent the player's interaction with these mechanics. They determine what each player does in response to the mechanics of the system, both individually and with each other. Sometimes game mechanics and game dynamics are used interchangeably, but they differ significantly.

And finally, the aesthetics of the system is how the player feels when interacting with the game. Game aesthetics can be viewed as the composite result of mechanics and dynamics, how they interact to create emotion.

Game mechanics

The mechanics of the game system consist of a series of tools that, when used properly, promise to bring meaningful response (aesthetics) to our players. For our purposes, we will focus on seven basic elements: points, levels, leaderboards, badges, challenges/tasks, onboarding, and engagement loops. We'll discuss each of these mechanics in detail, but we'll start with the heart of any game system - points.

Points

Points are important regardless of whether their accumulation is shared between players or even between designer and player. When you first think of a points system, you might immediately think of a goal in a sports game, exchangeable points in a video game, or bonus points awarded to players for successfully completing special in-game tasks.

Despite what your preconceived notion of the point is, it is sufficient to say that they are an absolute requirement for all gaming systems. As a designer, you absolutely must appreciate and follow every move a player makes, even if those points are only visible to you from your design console, not to the player. This way you can track how players interact with your system, design the results, and make changes and adjustments accordingly.

Point systems range from the obvious to the barely visible and serve a variety of purposes. As such, we would like to highlight a few types that you have undoubtedly encountered in your life.

Cash score

This is a number that tracks how much money you have in the bank. Considering how much we value money in our society, it's strange that we don't tell others our bank balance in casual conversation. Instead of breaking this social taboo, we let each other know our money score by what we wear, where we go on vacation, where our kids go to school, etc. Instead of literally shouting our scores, we signal them with status items. In this scoring model, signals tell the score, exact numbers do not.

Score in video games
Much more obvious is the score in almost every video game. The score is always on the screen, flashing in the corner of the screen, letting the player know how close or far they are from the next level, other players, and ultimately winning the game. Few real-life systems keep score as omnipresent as video games.

Score on social media

When Facebook was created, there was no apparent indication that the number of friends a user has would serve any function. Similarly, the number of followers or mentions on Twitter was never explicitly identified as a designated "score". But they are. Most users, if asked, can list how many friends and followers they have on a given social network. What's more, they can probably name who among their friends has an unusually high number of friends or followers. A sort of inventory is done simply because Facebook and Twitter have placed the "score" in a prominent place on the page.

Keeping the score

Google Orkut was a social network created by a Google employee. Most social network users in the world may not have heard of it. But in one place in the world, Orkut is the number 1 social network - it is Brazil.

When designing this page, Google introduced a simple leaderboard that listed the number of people in each country who had signed up for Orkut. When Brazil topped this leaderboard, something unexpected started to happen - Brazilians started to organise spontaneous registration parties on Orkut. The goal was to take over the top spot held by the United States. As a result, Orkut became the country's number one social network, easily beating Facebook, all thanks to the score and an innocent table.

Summary metrics

Any metric has the ability to become a type of score. Sometimes it is better to create a composite metric to convey complex data in a simple form. For example, a FICO score is a combination of a variety of different information - from average monthly credit card payments to accumulated debt over a lifetime. We could, of course, display different scores for each vector we want to measure. But by summarising the complexity of creditworthiness into a single number, anyone from a prospective landlord to a bank officer can derive meaning from it without needing a PhD in economics.

Weight Watchers is similarly creating metrics that reference a range of numbers - from body mass index and weight to daily calorie intake - to track the progress of its users as they strive for a healthier lifestyle. Likewise, Klout shows you and others where you stand on Twitter in terms of influence and relevance. It doesn't consider just one vector, but rather a series of numbers that are combined into one.

Point systems

In gamification, we can use one of the five point designs that form the basis of our experience. In some cases, your point system will be obvious, direct and highly motivating. In some designs, you will use four different types of points to achieve your goals. In others, the points will take a back seat to other mechanics, performing their duties in the background as the designer's workhorse.

No matter what you end up deciding, you will need to be well oriented to the basics and possibilities of the points system. Your points palette includes the following:

Experience points

Redeemable Points

Skill points

Karma points

Reputation points

Of the five types of point systems, experience points or XP are the most important. Unlike air miles, XP does not serve as any kind of currency within the system. They do, however, serve to track, rate, and manage your player.

Everything the player does within the system earns XP - and in general, XP never drops and cannot be exchanged. By assigning XP to every action in the system, you are aligning your behavioural goals with the player in the long term. In some systems, XP can expire - for example, monthly or annually - to create loops of goals. This pattern can be seen in retraining periods used in loyalty programs - and expiration can serve the important purpose of resetting the game to level the playing field.

More importantly, the number of XP points will never run out. The player earns them as long as he plays the game. That's the power of XP.

Redeemable points

The second point system consists of Redeemable Points (RP). RP, unlike XP, fluctuates - it goes up and down. Most people expect that these points can be exchanged for items within the system. They are earned and monetized much like a bank account balance or frequent flyer miles that we redeem for rewards.

RPs generally form the basis of the virtual economy and are often given names such as coins, dollars, cash, etc. As in any economy, capital flows need to be monitored, managed and adjusted to ensure that everything runs smoothly and that there is no massive inflation or deflation. In addition, there are significant legal and regulatory issues associated with the redeemable points.

Skill points

The third point system is called the skill point system. Skill points are assigned to specific actions in the game and are related to both XP and RP. It is a bonus set of points that allows the player to gain experience/rewards for activities in addition to the basic ones.

By assigning skill points to actions, we direct the player to complete some key alternative tasks and sub-goals. Classic examples of skill points can be found in Dungeons and Dragons and other similar games, here you have different skills such as magic and strength, and different scores for each. In a non-gaming context, you can assign a set of varied skill points on a photo-sharing website; players can earn some points for the quality of their photos and others for the quality of their comments. Depending on the circumstances, it may be worth separating them.

Karma points

Karma points are a unique system that is rarely seen in traditional games. The sole purpose of karma is to award points. This means that players do not gain any benefits from keeping their karma points, but only from sharing them.

A good example of this system is UserVoice. Using karma points, players vote for and against potential features they would like to see built. If a player's vote wins and those features are built, they get their karma points back to allocate to new activities, and so on. The only purpose of points in the UserVoice model is to give them away.

Reputation points

Finally, reputation points form the most complex point system. Whenever a system requires trust between two or more parties that cannot be explicitly guaranteed or managed, the reputation system is crucial. Its purpose is to act as a proxy for trust.

The reason why reputation systems are generally more complex lies in the way they are designed and used. In general, they need to cover a wide range of activities to be meaningful - and the design to date needs to take account of incentives and unintended consequences. Furthermore, because they are a proxy for trust, players will certainly try to trick the system. Integrity and consistency will be paramount.

How to use point systems

For starters, it is essential to string an XP system around your gamified system. It informs you and your players of activities that are more important.

In contrast, you should use redeemable points when you want to create a virtual economy. Virtual economies are most valuable when you want to motivate broad behaviour, large communities and/or use the economy to control behaviour. However, they also have unique challenges, such as legal and regulatory issues that are complex and rapidly changing.

Another challenge with redeemable points is the perception issue associated with them. If you announce a great redeemable points system, the first thing a player will do is go see what they can get. If what is offered is neither meaningful nor realistic, you may deprive the player of the system entirely. In other words, if the player looks at the buyout option and says, "Yeah, sure. I'm not going to win a free car for watching a video," then that's one player who may not believe it's worth staying in the system. Similarly, another player may see that he can get a free pizza for his points and think he doesn't like or want free pizza. In both of these cases, you may be in danger of losing the player.

Reputation points are complex but often necessary in the system. The biggest problem, however, is that they are easy to "play" with. TripAdvisor is a website that shares customer reviews of travel around the world. Due to the success of this site, it reportedly accounts for 30% of all hotel bookings. Therefore, hotels have an eminent interest in not only being favourably rated, but also in making sure that other hotels are not favourably rated. Although there are no official statistics, a cursory inspection of TripAdvisor immediately reveals a number of clearly "weedy" reviews.

Yelp, a site that allows users to rate local restaurants and entertainment venues, is also experiencing similar problems. The only way to know which reviews are real and which are fake is to read a lot of them. Neither of these sites has implemented a reputation system that matches the level of sophistication or value traded in the system.

Virtual economies

The power of the virtual economy lies in the fact that it allows the creator to bring in large amounts of money and control its outflow. Any student of macroeconomics may recall several communist and socialist countries that operated on this premise. In Cuba, a traveller can exchange any other currency for the convertible Cuban peso, but it can be almost impossible to exchange it back.

This is how most economies are conceived even in the virtual world. Zynga's Farmville, for example, has a completely one-way valve. The player can only put money into it, and there is no real-world reward to exchange; everything stays in the game.

In 2010, Zynga launched a well-known promotion with the companies 7-11 and Slurpee. Logically, you might think that a player could exchange credits for Farmville and pay for Slurpee in the store. However, the promotion was that for every Slurpee purchased, the player received free Farmville credits. In this case, the value of the virtual economy was higher than the value of the real-world reward for these players.

Virtual economy and secondary markets

Secondary markets have basically been the doom of the MMOG creators in which they have appeared. They were tolerated because their creators didn't count on them at all. In new game designs, however, the possibilities of secondary markets are often severely limited. Today's designers are trying to control as many aspects of the virtual economy as possible - and secondary markets are in conflict with that goal.

NOMINAL VALUES OF CURRENCIES

The perception of the value of a virtual currency can be closely linked to the currency in which players find themselves. For example, the equivalent of 1 dollar in the US is 1,000 Korean won (both of which will buy you approximately a soft drink from a street vendor). So when denominating virtual currency for Korea, it pays to multiply the base numbers by 1,000 versus the US. The new social games automatically renumber all player views by the player's country of origin and denominate everything in USD behind the scenes. The parallels with the real-world economy are not coincidental.

Dual economy

A well-functioning virtual economy will manage player demand with minimal difficulty. It allows a certain level of fluidity of the game. If a designer creates a promotion without the help of a virtual economy, he or she has to uncover new and difficult explanations every time. For example, "Tell three friends about us and you get a scratch card that gets you 20% off this or 30% off that, and if you tell four friends about us, we'll give you a basketball, and with ten friends...".

In a virtual economy, however, you need do no more to explain an action than tell the player, "Tell three friends and you get 200 points." In this case, the player needs to get two hundred points. Two hundred points does not need to be explained. The player already knows what he's going to get for it. That's why marketing is optimised.

In fact, Farmville demonstrates this very well in the so-called dual economy. It has created two currencies within the game system: money and coins. Conveniently, US dollars are converted into both. Each is used for different kinds of items in the game.

Dual currencies are more suitable for players who are already familiar with the nuances of the game. They are not necessarily suitable for newcomers. Thus, dual economy designs are unimportant at the beginning of the game. They can, however, reveal how fine a control a designer can have over player behaviour and demand if they have created a well-functioning virtual economy. By simply pulling these macroeconomic levers, you can guide players through the system.

For example, dual currency can allow you to set wildly different values for items within an economy while controlling the incoming money supply. In the real world, so-called dual currency systems have attempted to do this, for example in Cuba or China, where they try to keep things cheap for locals and expensive for foreigners. While this approach has rarely worked in the real world in the long term, it is coming to the fore in virtual economies - but with the previously mentioned objection about the complexity.

You may have chosen Explore, Comment, Join, Recommend and Give your opinion as your top five social actions and assigned the following values:

Explore: 1 point

Comment: 2 points

Join: 4 points

Give your opinion: 4 points

Recommend: 20 points

It can be reasonably explained that web exploration is the least beneficial of the five actions to the company's goals, even though it is important to the player's gaming experience. Therefore, you assign it the least value. On the other hand, player comments are worth twice as much because they can provide quite a bit of value to your community as a whole.

Joining in general is often an underappreciated action. However, getting players to log into your system is quite a challenge. So giving it actually four times the value it has to give to "explore" is actually the right idea. Having a player's email address and name is of great value - not to mention the new level of value they'll associate with your system when they join.

By giving "opinion" four times more value than exploration, you are making it clear that the social contribution of a player who supports your system is a very important activity - and perhaps even essential to your system. Obviously, recommending is one of the most extreme forms of player viral expression, so it naturally carries more value.

Levels

In most games, levels are an indicator of progress - although in this role they are no longer as exclusive as they once were. In the arcade game Miss Pac Man, for example, levels are clearly expressed by the colour of the ghosts, the layout of the board, and the type of fruit that loops around the board as the game progresses. Of course, designers of gamified experiences won't use traditional levels as found in video games, but understanding them can add a powerful tool to your design. Levels serve as a marker by which the player knows where they stand in the game experience over time.

Level design

In Miss Pac Man, the player immediately knows that the level has changed, because in addition to being told directly and seeing the colours change, the game has become more difficult and the rewards for rewarded behaviour rise. Meanwhile, while Miss Pac Man's avatar moves at the same pace, the ghosts move faster and the delineation of safety time zones is getting shorter.
In the game design, the difficulty of the levels is not linear. In other words, the first level is not half as difficult as the second level, which is not half as difficult as the third level, and so on. Instead, the difficulty increases in a curvilinear way. In the example of Miss Pac Man, the experienced player knows that after the third level, the ghosts slow down again and the demarcation of safety time zones increases. The game board may continue to hide additional complexity, but as with most level designs, the difficulty increases exponentially during each level and then decreases over time.

The complexity of each level of Angry Birds proved to be very appealing. With the help of strongly designed levels, the player almost fluently moves from one to another, gaining confidence and experience. However, at one of the much higher levels in the game - level 21 on the first board - the player encounters a level that is definitely more difficult than the previous one. In fact, it is so difficult that there is only one sequence of events that will get the player further. This is the first time in the game that the player is likely to notice that he has moved on to a new and more difficult level.

Some may find this move by the creators of Angry Birds controversial. Players who find the challenge too challenging will inevitably abandon the game. On the other hand, those who pass the level are more likely to feel that they have achieved something special and become part of a rare group. Clearing a level unlocks another board - so it is a significant achievement - and one that has made life difficult for the authors for quite some time.

Progressive difficulty

The average length of a Nintendo Donkey Kong arcade game is less than 1 minute. This is because the first level of Donkey Kong was incredibly difficult. Realistically, the game company in the 80s wanted its players to lose faster and thus put in more quarters sooner.

In today's gaming systems, we are interested in longer and stickier games. That's why today's designs start at the simplest levels and gradually move to more complex ones.

In PopCap's Plants vs. Zombies for iPhone, the player moves from one level to the next, with the game's difficulty increasing with each new level. A quick look at the first and tenth boards of the game illustrates how much more difficult the game becomes as the player progresses. The game board is visibly overcrowded with characters and obstacles.

In some systems, levels can define the complexity - or the leading element of the game, or they can serve as a passive marker that adds more depth and complexity to your system.

Either way, the best tips for level design are to make them logical or easy for players to understand, expandable so you can add levels as needed beyond the initial "boss level", and flexible. Finally, levels should be testable and upgradeable. Level balancing is as complex as the game design itself and should be tested and retested as you play.

Permanent levelling system

American Express has created an impressive level system using the verifiability of the credit card itself. In fact, most Americans, if asked, would probably figure out what you're talking about if you simply listed the rainbow colours of Amex credit cards: green, gold, platinum and black.

Funny thing about this list is that while green, gold and platinum are associated with money and precious metals, the fact that they ended up on "black" seems to be a surprising switch. After all, everyone clearly knows that gold is more valuable than silver. By using "black" as the highest grade, American Express has changed the meaning of the word black. Now "black" as the highest grade is as common as gold and platinum.

Precious metals

How do we know that gold is more valuable than silver and bronze? If you don't regularly follow the precious metals market, it's likely that you've been taught this fact by other levelling systems in your life - such as the Olympics - that use precious metals to design their levels.

University levels are similarly stamped with a clear ranking system both within institutions and for the population as a whole - bachelor's, master's, and doctoral or postdoctoral degrees, using language that clearly indicates levels of mastery at the university level. The Army and Scouting have arguably the two most sophisticated level systems of any institution. From badges and medals to titles like General and Eagle Scout, even uniforms indicate who is at what level of "game".

Progress bar

Progress bars have started to appear all over the internet. In most incarnations, they inform the player in the form of percentages how close they are to completing all the necessary information for enrolment. Basically, they are used to encourage new players to complete personal information on the site or to create a deeper core experience. Progression bars work hand-in-hand with the levels and serve as a guide for the player to progress. They work with percentages rather than naming completed levels. Note: the best progress bars never reach 100%.

Use of mataphor

Similar to Amex or Boyscouts, creatively describe the proposed levels for your gamified experience. Without using precious metals or gems, imagine what would be an interesting level system for your system.

Examples: use of metaphor

Let's say you're developing a gamified experience for a company that sells women's shoes. You decide it makes sense to name the levels after candy to evoke both playfulness and colour. The names you choose from the lowest level to the highest include - mint, cherry heart, marshmallow, chocolate, and truffle.

While these levels certainly sound like they might be attractive to your demographic, sometimes the problem with a literal system is that people lose track of where they are and confuse chocolate with a truffle, for example. Another thing to avoid is accumulating a list of levels that look "cute" if the cuteness isn't in your player's sincere expression.

Find your voice

One experience passed on from a potential client was eloquent. The client produced a type of financial market software that connected investors and trades. Their clients, investment bankers, already assigned value to animals such as bulls, bears and pigs. So they used them in a levelling system when developing a game experience for this demographic. But imagine trying to impress a 50-year-old investment banker with a picture of a cute little pig - it might not work. But that's exactly what they did - with damaging effect. Suffice to say, they had to remove the overly adorable figures and replace them with something more modest.

So while sometimes a designer may be drawn to use words that relate to a particular community - in the case of the bull, the bear and the pig, or even in the case of those chocolates and peppermints - perhaps the level shouldn't be represented by a literal or cartoon representation. Perhaps it would be more effective to use colour to represent, or even a representation of the word itself, depending on your audience. It's all about knowing your players.

People's charts

The purpose of the leaderboard is simple comparison. Not surprisingly, most people need no explanation when they encounter the leaderboard. By default, we see a ranked list with scores for each name and understand that we are looking at a ranking system.

In any 1980s arcade, a newcomer approaching a Galaga or Moon Patrol game cabinet would find a list of high scores on the screen - most of which would contain so many zeros that it would make a potential player uncomfortable. That's a terrible incentive to discourage playing the game! Even if the numerical scale is completely meaningless or messy, the player still feels that four million points is a lot and probably not easily achievable.

A leaderboard that does not discourage

Today's leaderboard has undergone a radical redesign since the boom days of pinball machines and neighbourhood game rooms. In the era of Facebook and social graphs, leaderboards are mostly a tool for creating social motivation, not demotivation.

They achieve this simply by taking a player and putting him right in the middle. It doesn't matter where he is in the leaderboards - whether he is ranked 81st or 200,000th - the player will see himself right in the middle of the leaderboard. Below him, he'll see his friends who are right on his heels, and above him, he'll see exactly how close he is to the next best score. And he'll know exactly what he needs to do to beat it.

However, if a player is actually in the top ten or twenty, then the leaderboard should directly reflect that. In the case of these players, the leaderboard should show their literal ranking, as this is likely to be meaningful to them.

Endless Leaderboard

In arcade games, with the exception of deleting the top scores on a daily basis, there aren't many ways to allow every player to be on the leaderboard of a given game forever. At some point, his score will be surpassed and he will drop out - or reach a certain number and sit there for weeks until someone finally beats him. In today's world, there are ways to control the leaderboards so that no player ever drops out or gets stuck.

Doodle Jump, a popular iPhone app, allows its players to watch the leaderboard split in different ways. The player can look at it locally, socially and globally. The local view shows him where he is compared to others in the system in his immediate proximity. In the social view, he can see where he stands among his friends and followers in the game. The global view allows him to do the same across the entire system.

There's no reason why players can't split their leaderboard any way they want. In fact, tracking a player's behaviour in the leaderboard will also give the designer information about the players. For example, a player who cares deeply about the leaderboard is likely to be a more competitive player and can be managed accordingly.

Leaderboards can also be displayed with a limited view available to the player. In a game with millions of players, this can be an important tool. Flight Control - the air traffic control game mentioned in the first chapter - has a leaderboard that shows other players playing at the same level, sorted by distance and time.

Other popular social network leaderboards include Klout, which ranks fans of the Sacramento Kings basketball team by their Klout score, for example - showing the social power of a given user on Twitter. Also, mobile app Yelp ranks a user's best weekly check-ins. Friends or royals -

the people with the highest scores in the game, much like mayors on Foursquare - can intersect the leaderboard.

Privacy and Leaderboards

Sometimes creating a leaderboard is not as obvious as it seems. When the items being compared are sensitive or difficult to quantify, leaderboards present a unique challenge. Given that the purpose of leaderboards is public comparison, how do you compare information that is best kept private?

For example, a gym has a strong interest in helping its users achieve a healthier lifestyle or meet their fitness goals. Therefore, asking a newcomer to come in, step on the scale, and have their weight compared to other gym members will likely deprive that gym of quite a few prospective members. Not only is public weight sharing potentially embarrassing for some people, but not everyone who joins your gym is there to lose weight. Some people want to train for a marathon, while others want to relax or even gain weight.

The first thing that might be clear is that more than one ladder may be needed to meet your gym goals. For a beginner, a leaderboard that lists his attendance can be a great introduction to the system. Runners might want to be part of a leaderboard where they compete against other gym members. And while bodybuilders might want to share their weight and publicly track their growth, people striving to lose weight might be less inclined to play the game if public humiliation is part of the game. Further, there is the possibility that some of these leaderboards could trigger unhealthy outcomes if players are trying to win.

Creating leaderboards using sensitive or private information is challenging, but not impossible. Abstract scoring systems can ensure that each player maintains a program that is healthy for them while evaluating their achievements in a public leaderboard. Ultimately, the designer will need to keep their goals in mind and maintain awareness of their overall goals - and take some responsibility for the power of the leaderboard.

In some cases, for truly competitive people, a straight leaderboard can be a powerful motivational tool. For most explorers, socializers, and many successful people, it can be both positive and negative. Consider the motivations of your players and create a social leaderboard: it's a win-win.

A Survey of Social Network Analysis

Jerry Chun-Wei Lin[1] and Panus Jan[2]
[1]Western Norway University of Applied Sciences, Bergen, Norway
Email: jerrylin@iee.org
[2]University of Pardubice, Pardubice, Czech Republic
Email: Jan.Panus@upce.cz

1. Introduction

The nature of the world economy has changed, and the techniques used in international business have been affected by globalisation. Not only large corporations, but also small and medium-sized businesses are making efforts to sell their products and services worldwide. These companies are considering how to expand their production and service offerings to countries where it is profitable to redistribute their own activities. These activities are being relocated to countries that can offer them more favourable conditions. Not only have international trade flows become more dynamic in recent years, but so have technological paradigms. Products are manufactured on a global scale through a variety of global supply chains. Not only the operations of politicians, but also the activities of economically active organisations are causing borders between countries to disappear, while at the same time increasing competitiveness in international markets. Economies in transition benefit from this impetus to their development.

It is more than obvious that trade occurs more frequently within a country than between countries (Helliwell, 2000; McCallum, 1995). Eaton and Kortum (Eaton and Kortum, 2002) estimate that regions that do not have geographic borders can trade five times more than regions that do have borders. Many formal barriers to trade, such as inadequate enforcement of international treaties with the help of governments or lack of knowledge about the many opportunities for international trade, are often cited as explanations for the lack of trade among nations. Building international business and social networks can be an effective way to address these barriers. Subsequently, researchers can assist in determining strategies to circumvent these barriers. Research results can be used as records and can quantify the presence of such barriers.

While most of the research on the impact of domestic networks on international trade is motivated by the view that domestic networks are an informal barrier to trade in which network members collude to increase their market power by restricting foreign competition, transnational networks are the focus of research as a means of overcoming trade barriers. Domestic networks, on the other hand, are the subject of much of this research. The impact of domestic networks on the structure of international trade is also the subject of some research, which can be found in another area.

The purpose of this study is to draw attention to the potential of analysing international trade between nations in the context of social network analysis and to highlight some of the ways in which social media can use social networks to improve people's perceptions of companies. The next sections of this work provide various definitions of economic networks and international trade. These definitions will be included in this study. The field of international trade analysis is covered in a large number of publications, but social network analysis as a method for analysing international trade is addressed in only a small portion of these articles. In this presentation, we will address the potential applications of social network analysis in international trade. In order to perform an analysis of certain fundamental aspects of international trade, we will use a random graph model.

2. Social networks analysis: methods

The notion of an individual's popularity within the surrounding group of individuals who are similar to him/her is the fundamental principle in evaluating any statistic. This person is now in the spotlight of the room. According to (Moreno, 1934), who was one of the first researchers to study the effects of certain students in school, students made friends based on the choices they made. When the concept of social network was originally introduced, it was sociologists who formally defined its meaning (Bavelas, 1950). The idea of centrality is actually quite simple: it refers to the point (a person) that is at the centre of other points (a group of individuals) that have some form of influence. The idea of points, often referred to as the degree of points in the graph, was first introduced by Nieminen (Nieminen, 1974). Graph is the word used in certain mathematical constructions studied under the heading of graph theory (Gross & Yellen, 2003; West, 2001). Vertices, also called nodes or points, are the building blocks of graphs and are connected by edges (also called links or lines). A diagram of a graph can be seen in Figure 1. Graphs can be either undirected, meaning that the lines are connected symmetrically, or directed, meaning that the lines are connected asymmetrically and have a specific direction in which to move.

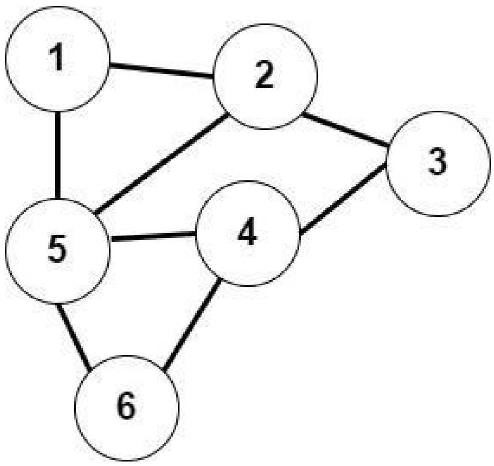


Figure 1: An example graph

An intuitive concept of how effectively this point is established in the network correlates with the number of degrees measured at this location. Measures based on degrees can be used for graphs that are either directed or undirected. The work of Knoke and Burt (Knoke & Burt, 1983) introduces the idea of in-centrality and out-centrality at different points on the graph.

Figure 2 provides a good illustration of degree centrality. Since it has a value of degree 4, which measures the number of connections to other places, point A can be considered a regional centre . The value of degree for points B through F is only 1. This method of measurement has a number of disadvantages, one of which is that it is difficult to compare the results with those of other charts. If you compare the same graphs with the same number of points, it provides relevant information. The size of a point is proportional to the overall size of the graph. A relative measure of local centrality was introduced in (Freeman, 1978). The total number of points in the graph determines how to interpret this relative value. It is not the same to have a point with a degree of 10 in a graph with 20 points as to have a central point with the same degree in a graph with 50 points. Point A in Figure 2 has a relative degree calculated as: 4 (degree) / 4 (number of total points minus itself) - 1.0 (Freeman, 1978, 1980) also offers overall centrality, often called global centrality. This idea, which can also be called proximity (Bavelas, 1950), or it contains certain concepts comparable to those of global centrality.

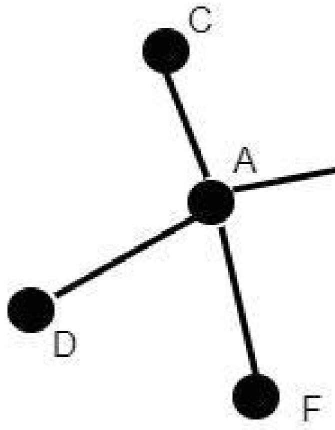


Figure 2: An example of degree centrality

Computing the Closure Centrality of a node involves determining the length of the route that represents the shortest distance between that node and any other node in the network. This can be thought of as the distance between points in the graph, and it can be thought of as two points being connected if there are more lines connecting the two points themselves. Sabidussi (Sabidussi, 1966) proposed the concept of proximity as the sum of geodesic distances between points on a graph that was not oriented in any particular direction. There are many different areas of study that can be described using proximity and centrality as examples of the network. Okamoto et al. (Okamoto, Chen, & Li, 2008) present some concepts of emotional proximity to a group of women in the social network. Wu et al. (Wu, DiMicco, & Millen, 2010) analyse the behaviour on certain business social network sites to find out which interaction patterns signal proximity. Roberts and Dunbar (Roberts & Dunbar, 2011) provide some concepts of emotional closeness between a group of women in a social network.

Another theory of centrality that Freeman (Freeman, 1978) calls betweenness is also included in his work. This measure indicates how often a point serves as a bridge between two other points (Jones, Ma, & McNally, 2021). The bridge can be thought of as the point in the diagram that makes connections between various other points. As can be seen in Figure 3, point B serves as a connection between two separate subgraphs. It happens that nodes with a low degree of centrality nevertheless play an important role in the network by connecting the rest of the nodes. Since these points have the ability to control the connections and information flow between other nodes, they are often referred to as brokers or gatekeepers. Burt (Burt, 1992) uses the term "trust points" to convey the concept of "structural weaknesses" This is usually the case when there is only one point in the network that has high interconnectedness, as shown in Figure 3. Relying on this argument is also fraught with significant dangers. The person has the option of leaving the organisation or not connecting with another group, both of which result in the flow of information being interrupted. It is possible to represent structural holes as two locations separated by a distance of two points, while other points are separated by a longer distance. In the diagram shown, point C cannot contact any of the other components of the network without going through point A first.

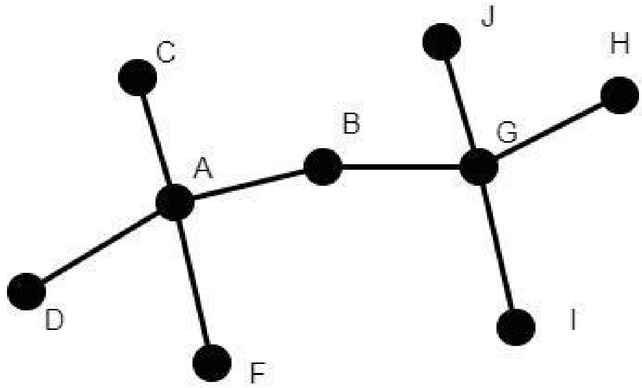


Figure 3: An example of betweenness

There are a considerable number of academic writings that focus on betweenness as a central concept for the purpose of analysis. Kourtellis et al. (Kourtellis, Alahakoon, Simha, & Tripathi, 2013) offer an alternative method to discover nodes with high betweenness. They use a randomised centrality measurement technique that considers k-path centrality (Alahakoon, Tripathi, Kourtellis, Simha, & Iamnitchi, 2011). An interesting and illustrative perspective on betweenness (Everett & Borgatti, 2005). They perform an analysis of what they call the effective ego network, which is described as a network consisting of individual points (egos) that are connected to other groups. Brandes (Brandes, 2001) presents several modifications of existing techniques to allow faster computations of betweenness and centrality in large networks. Goh et al. (Goh, Oh, Kahng, and Kim, 2003) establish a link between the neural network and the social network and investigate the relationships between betweenness centrality and disassortative networks and neural networks.

3. Social networks analysis: random graph model

The creation of models and simulations of real-world problems is called random graph modelling (Albert, Jeong, & Barabási, 1999; Erds & Rényi, 1966; Molloy & Reed, 1995). Random graph modelling is a general term. It is a branch of mathematics that bridges the gap between graph theory and probability theory and refers to the distribution of probabilities over graphs. Random graphs are used to answer questions about the properties of graphs. They are often used as an explanation of complicated networks. It is necessary to model these networks to understand their complexity. When using a model based on a random graph, it is important to follow these procedures. It is important to consider each network connection as a separate random variable. The accumulation of such data points can be interpreted as a connection with a certain degree of probability. It is difficult to describe how such a network works and what principles are used to establish the connections. It is more accurate to say that we have limited understanding of both the networks themselves and the process of building networks of this type. We can explain that our model is not able to make an absolutely accurate deterministic prediction, and we can also say that the result contains some noise that we cannot explain. It is the responsibility of the individual to establish new relationships that do not depend on those established in the past. We need to make a working hypothesis for the construction of a network by certain particles and then include them in the model. The construction of the network is reflected in each of the parameters of the network. Each potential network connection can be viewed as one of many different network configurations. The model reflects the probability distribution of a random graph, and these configurations point to structural aspects of interest.

Researchers advise keeping settings as simple as possible. This complexity is reduced by applying homogeneity and other constraints. If the researcher restricts the number of parameters, the model can be defined in a more effective way. The researcher will treat some of the parameters as equivalents in order to unify or relate the remaining parameters in various ways. The construction of the model and its interpretation, including its parameters. The attention paid to the development and understanding of the model is at the heart of the modelling process. On the other hand, this strategy often requires inference from the four principles previously discussed. When the model structure is sophisticated, as is the case in most real-world situations, this last phase is extremely difficult to successfully accomplish. In the case of parameter estimates, as well as an estimate of the uncertainty of the model, researchers often take advantage of what statistical models for networks have to offer. If we take a certain number of nodes, denoted by N, and connect them such that each pair of nodes, denoted by i, j, has a connection with an independent probability of p, we create a random graph. However, if we want to study models that approximate the actual basis, we have to accept the fact that such a simplified model has certain weaknesses. One of them is the distribution of degrees on the graph, which must be calculated in the same way as in the real world. In most cases, it is claimed that the probability of a graph can be determined just by counting the subgraphs of the graph. Consider a node in a random graph. Using binomial division, determine the probability that p is connected to each of the N-1 other nodes in the graph. This yields the probability that p_k is associated with each node based on its position in the network.

International trade provides many business opportunities, greatly increases employment, international transport, brings new impetus to solving different problems, it helps to introduce new methods in various areas of economics or politics. International trade can be taken as one of the main pillars of the world's economic fabric. The great benefit is the creation and dissemination of different business databases, both at regional and global level. Transnational networks can facilitate interconnection through commissions of marketing operations that let potential traders know that there are customers who are interested in the product in another country (Dreyfus, Dreyfus, & Athanasiou, 2000). Within a given market, these networks can help find a suitable distributor of goods for specific customers (Weidenbaum & Hughes, 1996).

It is common practice to use international trade as the basic unit for a country's descriptive statistics, including things like import quotas, pricing, comparative advantage, exchange rates, and so on. There are a variety of statistical methods that can be used to analyse international trade. One of these is the relational method, where the focus of the study is on trade flows between nations. Therefore, I propose to use the study of social networks as a method of analysis, a strategy that is very pertinent. Social network analysis can be particularly useful for this study because it maps the overall structure and growth of international trade, which is more than just a collection of bilateral interactions. In today's world, international trade is seen as more than just a collection of bilateral relationships. The results of empirical studies show the positive influence that international trade has on the conditions created for trade between groups operating across borders and for immigration. Immigrants know the characteristics and qualities of consumers and sellers in their home countries, and they bring this information with them to the new countries to which they move. However, it can be difficult to assess the extent to which the influence of transnational cooperation would work by providing information in the marketplace or by using official information. This is due to the nature of the phenomenon under study. Sometimes a customer's demand for products from his own country is obvious but has no influence on a network because he is not part of a network (Gould & Fernandez, 1989). For the years 1970-1986, Gould calculates various import and export equations and the impact that immigrants had on bilateral trade between the United States and its trading partners.

The results of this study point to a number of possible implications for the economic effectiveness of transnational networks that provide information about lucrative business prospects (Rauch & Casella, 2003). They use a model that considers the following: The producer must match his requirements to determine whether or not such an arrangement is acceptable; if so, then it is feasible to hire an internationally immobile labour force; and finally, the production is carried out. Based on his previous experience in his own country, the manufacturer can match his requirements (i.e., what he needs to produce and with what resources he can do it) with the resources available to him. The typical requirements and information needed in other countries are not as well known to manufacturers from other countries, leading to complications. Thus, consensus at the international level could help shift the demand for labour in the form of services to manufacturers from countries where this type of labour is scarce to countries where there is an adequate supply. In trying to define and understand networks on a global scale, using social networks as an analytical tool is an effective approach. There are a variety of challenges in transnational collaboration, many of which require building relationships with other producers or workers who know the environment and can serve as an appropriate tool to determine how to effectively utilise resources in those countries. The exponential random graph model (ERGM), explained in the next chapter, is therefore a suitable tool for analysing and modelling existing relationships. The form of the exponential random graph model, which is a subset of the random graph model, is as follows:

$$P_r(X = x) = \left(\frac{1}{k}\right) exp\left[\sum_{A} \mu_A z_A(x)\right], \tag{1}$$

where X_{ij} is a random graph showing how the actors are connected. Next, X represents the element matrix n and x represents the matrix of network connections achieved. The letter A then stands for an additional network of configuration types. A set of dependent variables of the model is denoted by the notation $Z_A(x)$, which states that any given collection of facts The chance of constructing a particular network is affected by a computation performed with x. The coefficient of A is the coefficient of the network. The coefficient μ_A is a representation of an unknown parameter; in the supervised network model, this parameter evaluates and conveys the influence of network statistics. The value of the coefficient k indicates the number of different counters that can occur in a network with n different components.

Originally, parameters in ERGM were calculated using pseudosimilarity (Strauss & Ikeda, 1990), but this technique is generally not trustworthy. Instead, it is more acceptable to use the Markov chain of Monte Carlo as a similarity estimator (Geyer & Thompson, 1992). The distribution of a random graph is simulated using Monte Carlo, using the initial values of the parameters. This process is continued until we reach the adjusted values, which we then compare with the simulated distribution of the graph provided using the collected data (Snijders, 2002). The advantage of this method is that it gives us an estimate comparable to maximum likelihood estimation and a number of standard errors that can be relied upon even when there are an unlimited number of possible distributions of the network configuration (Wasserman & Robins, 2005). The development of a number of other network specifications, collectively referred to as social circle dependence (Pattison & Robins, 2002), has led to a significant reduction in the number of cases where a high number of triads often occur due to inadequate model specification (Hunter & Handcock, 2006; Robins, Pattison, & Wang, 2009). The ERGM is described as an exponential version of the log-likelihood function, and this particular function typically has a surface that is concave everywhere in the world.

There are a number of works that explore the potential applications of ERGM for modelling the connections between specific elements (Kim, Howard, Cox Pahnke, & Boeker, 2016). It is possible to study the effects of independent changes on the characteristics of different trades, as well as the dyadic coefficients and structural effects of these trades. These characteristics are highlighted by the authors as significant in the process of establishing new links in international trade. When two things have similar characteristics, they are more likely to be related in some way. The research theme followed by the research project is an essential component that plays an important role in both social network analysis and ERGM analysis. It is not the distinguishing characteristics of the things studied that can be a source of great difficulty, but the solution lies in the nature of the connections that exist between the things studied. Moreover, ERGM integrates substantive and relational factors, as well as quantitative and qualitative analysis, which places a great responsibility on the shoulders of the researcher. After that, all the concepts related to the research topic must be well defined and the results of the study must be accurately interpreted. Unlike other metrics, which can usually be interpreted based on common sense, some measurements can be directly translated into import and export values, with no room for interpretive debate.

Closeness centrality can be used to determine the distance between a node and the other nodes in a social network, while betweenness centrality can be used to determine the frequency of nodes along the shortest path through the entire network. Both concepts can be used for further analysis of social networks. However, when applied to international trade, these numbers are notoriously difficult to interpret, although some authors choose to do so. One option is to examine the adjacency matrix, but this can only be done for a certain number of countries. These countries are then subjected to an appropriate analysis that allows the extent of the links between them to change over time. When detecting links that are comparable to those of other entities in the network, the so-called structural equivalence analysis is an acceptable method of investigation. This method can be used effectively for categorising nations in a variety of different business network activities.

The ERGM approach is an effective tool that can be used in situations involving the expansion and development of business opportunity strategies or the acquisition of new customers. It is possible to analyse the components under study in more detail and provide a better description of the interdependencies in the networks. The results can then be used to explore the relationship using rigorous scientific and empirical evidence to build models (Colquitt & Zapata-Phelan, 2007). The influence of organisational resources on alliance formation is significant in that multi-source organisations have a greater number of opportunities to create linkages and their higher level allows for the need to collaborate through alliances to be waived. Several studies have pointed out the importance of making use of microtransactions in interregional and transnational cooperation. Again, the ERGM technique is best suited to create models of existing links between individuals and their interconnections to establish links between organisations and companies (Gulati & Westphal, 1999; Rosenkopf, Metiu, & George, 2001). Individuals who have worked in multinational corporations in the past start new firms. As these new firms are linked to employees of existing corporations, the large corporations may gain additional motivation to promote their collaboration.

An acceptable method that can be used in a variety of subfields of international trade research and modelling is called ERGM. It is suitable not only for modelling production and supply networks, but also for planning and analysing strategic studies in the field of international trade. Given that people, organisations, and the community as a whole are increasingly interconnected, methods need to be developed to gain a deeper understanding of the overall architecture of the network. ERGM thus extends the traditional foundations of social network analysis and takes them in a new direction to gain a deeper understanding of the origins of social networks.

4. Social networks analysis: Artificial intelligence

With the exponential growth of social network data, artificial intelligence (AI), machine learning, and data mining technologies are widely used in social network analysis and have various types of tasks, applications, and properties that researchers focus on.

4.1 Types of analysis tasks

As mentioned earlier, social network data can be represented and analysed in graphs. Tan et al. classify social network graph analysis tasks into four main types, namely node classification, node clustering, link prediction, and anomaly detection (Tan et al., 2013), as shown in Figure 4. In social networks, users are often tagged with various attributes such as gender, education, location, beliefs, and interests. However, collecting and labelling users is time-consuming and costly, and there are few users with labelled data. Therefore, node classification tasks aim to predict (classify) the unlabeled users (nodes). As the old saying goes, like goes with like. The labelling of a user can be predicted by the information about its connections and neighbours .

What if we do not have some properties at all? For example, as elections approach, candidates may be interested in analysing the party orientation of users on social media. However, there may be no labelling of party orientation for training machine learning data, and the task becomes unsupervised learning of clustering. Node clustering tasks in social networks aim to divide users (nodes) into groups where users in the same group are similar to each other and less similar than users in other groups.

As for link prediction tasks, a link represents the connection between nodes. It may be that connections missing between users need to be repaired, or that the connection of new users starts. For example, if a user is in the process of signing up for a new account, the system may suggest other users the user may know and subscribe to channels the user may be interested in. The last task, anomaly detection, is receiving a lot of attention in social network research and aims to detect malicious and negative events such as porn, spam, smuggling, violence, harassment, hate speech, fraud, and terrorism. The malicious events can traditionally only be reported by humans and can be automatically detected by machine learning and data mining algorithms. Artificial intelligence enables social network analysis and management to make accurate predictions and respond quickly.

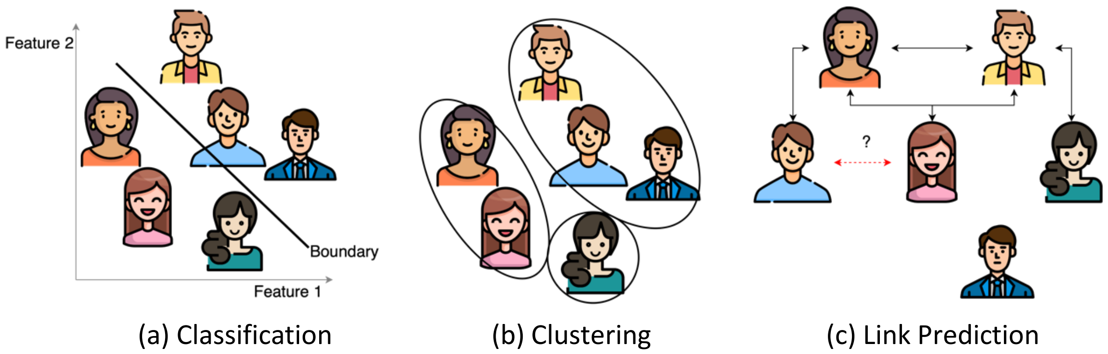


Figure 4: Examples of classification, clustering, and link prediction

4.2 AI-Based Applications

Among the above learning tasks, anomaly detection has various applications in daily life. In social network management, anomaly detection includes malicious user identification (see Figure 5), fake news, sentiment detection (depression, hatred, etc.), and personality trait detection. When identifying malicious users, they can also be identified by login network IP address, device IP address, and physical address. For example, a user who logs in from different device IPs, network IPs, or physical addresses in a short period of time could be misappropriated. On the other hand, if a network IP address attempts to log in to multiple accounts, it could be operated by criminals. In addition, connections to criminals can also provide information to detect potential illegal events.

As social platforms are part of people's daily life, many fake and fraudulent news are spread in social networks, which are harmful to both physical and mental health (Naeem et al., 2020). As more and more fake news is spread on social networks, the demand for fake news detection analysis is very high. Researchers (Zhou & Zafarani, 2018) examined fake news research and pointed out that fake news harms democracy, justice, and public trust. For example, during a pandemic, fake news and information have a negative impact on disease prevention, control, and cure. Natural language processing techniques can be applied to text-based information to detect potential Fake News. On the other hand, the connections between Fake News broadcasters on graph-based information could also help in the analysis of Fake News.

Sentiment recognition is an important task for social media management. Semantics that need great attention include depression, denigration, hatred, and racism, to which well-known social media companies (Facebook, Twitter) pay much attention. Spreading depression and denigrating information could lead to suicide, and spreading hate and racist information could lead to violent behaviour . Therefore, social media should recognise, restrict, or provide a warning about the relevant information.

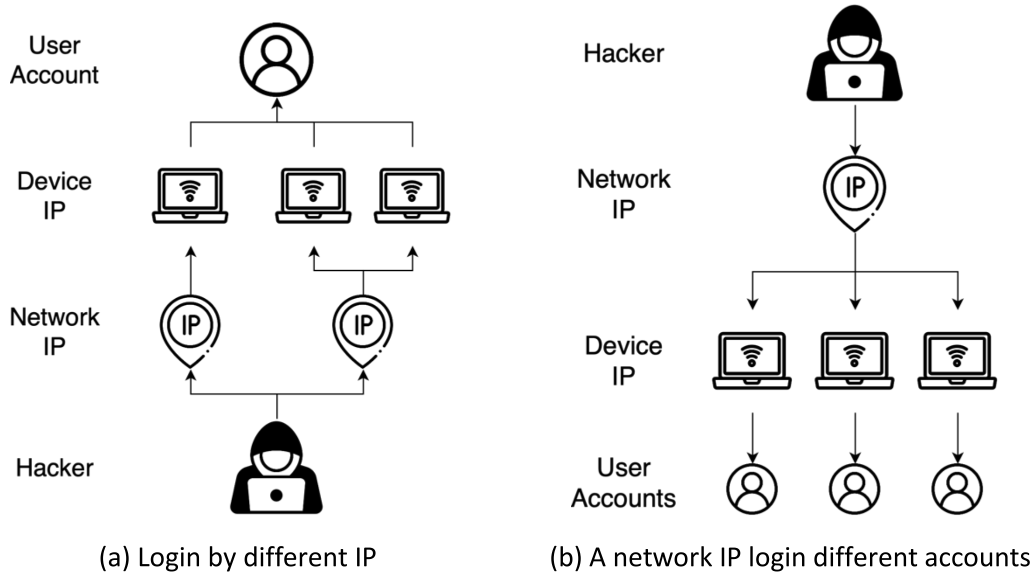


Figure 5: Examples of malicious user identification

4.3 AI Techniques for Social Network Analysis

Data mining aims to efficiently extract interesting and meaningful patterns from huge amounts of data suitable for research topics in social networks. Numerous data mining algorithms have been proposed for relational data (Jensen & Neville, 2003) and applied to social network topics such as goal guidance (Yang et al., 2006), suicide prevention (Lopez-Castroman et al. 2020), anomaly detection (Bindu & Thilagam, 2016), and sentiment detection (Zucco et al., 2020). Take sentiment detection in Section 3-2 as an example. In 2020, 'covid' and 'coronavirus' became keywords in social media (Nemes & Kiss, 2021), which greatly affected the sentiment of users and also physical and mental health. The techniques of text analysis and data mining can be used to analyse and present the trend of public attention in social media (Hou et al., 2021). Sharma and Sharma (Sharma & Sharma, 2021) also proposed a signature-based unsupervised sentiment analyser to analyse people's depression and suicidal tendencies on social media.

Deep Learning is another powerful group of algorithms for social network analysis. The first task is data processing, which encodes the large amount of social network data into low-dimensional features for deep learning algorithms (Tan et al., 2019). Deep-learning-based representation algorithms include embedding, autoencoder, and graph convolutional approaches. Embedding is a transformation approach in which features are mapped into the corresponding vector through lookup tables. This can be a deterministic or a dynamic matrix learned from an attributive or a heterogeneous network representation (Tan et al., 2019). In the autoencoder approach, the encoder converts the input features into a low-dimensional vector and the decoder outputs the recovered features. When the difference (loss) between the input and output features is small, the representation of the low-dimensional vector can well represent the high-dimensional vectors (since it can be highly recovered by the decoder network). In the graph convolutional approaches, convolutional neural networks (CNNs) are used to extract the neighbourhood information. Advanced inductive training and graph attention techniques can also be incorporated into graph convolution approaches to better represent the features.

According to the feature representation, Deep Learning can predict the network information in various tasks. Take the identification of malicious users mentioned in Section 3-2 as an example. The network, which consists of the network IP, device IP, and physical address, could be analysed by the Graph Attention Network. On the other hand, the time variant of the features could be detected by deep recurrent neural networks. After analysing the network features, the malicious user could be identified by the deep neural network (Ye et al., 2022). From the other perspective, user identification could also focus on linking user identities, i.e., linking the same user on different social platforms, and can be fulfilled by data of profiles, generated contents, and behaviours. Zhou et al. (Zhou et al., 2018) proposed a semi-supervised learning method through coded vector representation with local and global network structures.

4.4 Critical Properties for Artificial Intelligence Methods in Social Network Analysis

The basic but critical properties of AI methods in social networks include dynamic networks, heterogeneous data, scalability, and interpretability, which leads to various researches to improve the properties of AI methods (Tan et al., 2019). Since social networks are a rapidly changing environment, dynamic network is a critical property for AI-based methods in social networks. The highly dynamic nature of nodes, relationships (edges), and embedded patterns can change over time. For example, new social media users will join the network and form new nodes, and friendships between nodes will also be updated. These changes may affect the performance of AI models in the existing network. Therefore, regular performance update and monitoring is essential for AI models in social networks. The current research mainly focused on the basic changes of deleting and adding edges (Li et al., 2017). However, the social network environment is more complicated, such as version changes, additional tools (new dimensions of nodes and edges), and recommendation algorithms. Therefore, the dynamic network is an important but still open question for AI-based methods in social networks.

Heterogeneous data is also a critical issue for social network analysis. An AI-based model that works well on homogeneous data may not work well on heterogeneous data. For example, a model developed and trained for Facebook data may not be useful for Twitter data because the two data sets have different data types, user characteristics, and patterns. The lack of data input can have a large impact on AI-based predictions and decisions. Xie et al. (Xie et al., 2021) summarise that the heterogeneous information contains distinguishing nodes and relationships (edges), and heterogeneous networks aim to learn a representation of each (different) node to reflect the semantic information. Meta-path-based proximity is an emerging method for embedding heterogeneous information networks (Wang et al., 2020); however, related research is still at an early stage.

When it comes to scalability, the efficiency of the methods should also be considered, even though deep-learning-based methods have shown excellent performance in feature representation and prediction. Social network data has high complexity in the form of different data types (numbers, text, and numeric data) and complicated associations (friendship, sharing, and comments). Moreover, huge amounts of data are generated in real time, such as Twitter and Facebook. The question of how to achieve scalability and efficiency of deep learning-based methods is a major concern in social network research. One possible technique for scalability is parallelisation, either with Graphics Processing Units (GPUs), parallel computing, or distributed learning.

Despite the robust predictability of deep learning-based methods, one of the drawbacks is the lack of interpretability. Take the feature representation in Section 3-3 as an example. The embedded vectors do not have intuitive and explainable meaning, but complex matrix transformations in latent space (Tan et al., 2013). Interpretability could ensure the correct and meaningful learning patterns of the models and also provide confidence for the decision-making process, which has attracted much attention in research (Du et al., 2019).

5. Social networks analysis: Using social media as a customer interaction

This new platform of social media and social networking represents one of the most remarkable technological advances of the new century, and this social media has allowed users to create and share content on a variety of different platforms. Businesses have wisely adapted to this new trend, giving many companies the opportunity to use social networking platforms to successfully engage with consumers. According to Mangold and Faulds (Mangold and Faulds, 2009), businesses are attracted to social media because it allows them to reach millions of users. Currently, the vast majority of companies use social media marketing to get as much of the online population talking positively about their products as possible. The emergence of Internet-based social networks enables communication with thousands of people about the products (and services) and the companies that offer them. As a result, social media has become an important means for companies to communicate with their markets or target markets (Kurtz & Boone, 2012). Businesses looking to capitalise on the prevailing trends have more options today than ever before with the expansion of various social networking tools and websites. With the help of these tools and websites, companies now have access to a new communication and engagement channel. According to Kaplan and Haenlein (Kaplan and Haenlein, 2010), decision makers and consultants alike are looking for methods on how companies can turn apps such as Wikipedia, YouTube, Facebook, Second Life, and Twitter into lucrative ventures for the company (Kaplan & Haenlein, 2010). Businesses use social media as a means of networking and interaction as a seamless means of direct interaction through word-of-mouth and is great for reaching target audiences and has several popular applications such as blogs, instant messaging, and widgets, which are on-screen tools for interaction. This is in contrast to traditional forms of networking, where people meet in person at events or conferences and exchange business cards. The use of social media platforms by companies for conventional and core marketing, as well as for customer contact, is becoming increasingly popular, as social networks offer added value in the form of a tool for consumer interaction and networking.

Vollmer and Precourt (Vollmer & Precourt, 2008) point out that consumers turn to different types of social media for product information and purchase decisions, making the need for social media as a customer interaction tool even more relevant. Not only are companies turning to social media, but so are consumers (Vollmer & Precourt, 2008). For this reason, businesses are expected to make the best use of social media to communicate with their customers or consumers and ensure the availability of information when it is needed. According to Muniz and O'guinn (Muniz and O'guinn, 2001), many companies are already using social media and networking sites to either help build their brands or to pursue marketing goals. This has also been noted in the academic literature. The ability to communicate with an ever-increasing number of customers is enabled by social media, which is a beneficial potential for telecommunication networks. The purpose of this study is to analyse the use of social media by telecommunication companies in developing countries as a tool for interacting with their customers, with Ghana serving as the centre of this research.

According to Kaplan and Haenlein (2010), 75% of Internet users participated in social media in the second quarter of 2008, either by joining social networks, reading blogs, or posting reviews on shopping websites. This number represents a 56% increase from 2007 (Kaplan & Haenlein, 2010). Someone once told me that if you type the name of a well-known brand into a search engine, there is a good chance that a company website will come up. This means that companies also make sure that the necessary information about their goods and services is accessible to both their existing customers and those who might become customers in the future. According to Gallaugher and Ransbotham (Gallaugher and Ransbotham, 2010), social media is fundamentally changing the contact between companies and their customers. According to Sashi (Sashi, 2012), the interactive features of Web 2.0 have led to an explosion of interest in customer engagement. Moreover, the opportunities that social media offers for building strong relationships with customers seem to excite practitioners in a variety of industries around the world. The use of social media represents a great opportunity for companies to improve their customer relationships and, in turn, their revenue, operational efficiency, and cost savings. Even though the use of social media is associated with all these benefits, companies can only lose if they do not make the most of the advantages that the use of social media offers. According to Baird and Parasnis, the use of social media (networks) as a platform for customer engagement will be unsuccessful if social customer relationship management (CRM) strategies are not redesigned (Baird & Parasnis, 2011). If companies want to get the most out of social (media) networks as a form of consumer contact, they must redesign their legacy approaches to customer interactions to fit the unique requirements of the markets in which they operate.

It is well known that the use of social media as a method of engaging consumers is widespread in developed countries. However, there is relatively little information in the published literature about the prevalence of similar use in developing countries. As more and more people in developing countries have Internet access, businesses in these countries have more opportunities than ever before to engage with their customers through social media. Therefore, it is important to examine the extent to which companies in developing countries are taking advantage of the opportunities presented by the aforementioned advances in their day-to-day business operations. In line with the issues discussed earlier, this study will examine the extent to which telecommunications companies in developing countries (particularly in Ghana) are using social networking platforms (such as Twitter, Facebook, LinkedIn, my space, flair, etc.) as a tool to interact with their customers. The study will focus specifically on Ghana.

To achieve this overarching goal, the research will have the following specific objectives: (1) gathering information about the types of customer interactions for which social media are used, and (2) analysing the information gathered to identify areas for improvement. We will offer some suggestions on how the media covering telecommunications can use social media more effectively. It is expected that companies in developing countries will adopt new strategies and means to interact with their customers and clients, given the increasing use of the Internet worldwide and the increasing penetration of the Internet and cell phones in developing countries. This is due to the fact that developing countries are increasingly using the Internet and cell phones. It is noted that social networking interaction is the current trend for businesses to engage with their clients regardless of their location. This is due to the extensive use of social networking sites and platforms that is now being seen in both developed and developing countries. Although many articles have been published on how social networks and social media are being used to communicate with customers in developed countries, there is very little information on developing countries.

This means that the number of customers that need to be served by these telecommunications companies is increasing every day, making it extremely important to interact with a larger number of people. Mobile communications services are becoming increasingly popular in Africa and many other developing countries. As communications companies and their customers (and potential customers) have access to the Internet, there is a good chance that social media can be used as a means to interact with customers and provide solutions and answers to their questions and problems. This is a possibility that is highly likely to occur. For this reason, it is important to study the extent to which this is happening, especially in the context of developing countries. For this reason, the research is of great importance because it has the potential to provide recommendations for better activities based on the results. These ideas could be based on the results of the study. To provide answers to the research questions, this study utilises a qualitative case study methodology. Specifically, the researchers will collect data and conduct analysis using selected telecommunications companies. The interviews will provide qualitative data, while the statistics found on the websites will help provide quantitative data. The interviews will provide qualitative data. The results of the two organisations will then be compared, which will facilitate the formulation of proposals to improve the quality of interaction with customers through the use of social networks in developing countries.

In marketing literature, social customer relationship management (CRM) is defined as the integration of customer-facing activities such as processes, systems, and technologies with new social media applications to engage customers in collaborative conversations and improve customer relationships (Trainor, Andzulis, Rapp, & Agnihotri, 2014). According to a study by Malthouse (Haenlein, Skiera, Wege, and Zhang, 2013), the proliferation of social media poses a threat to the traditional customer relationship management (CRM). Traditional CRM is undergoing a modest but steady transformation triggered by the increasing use of social media as a channel for consumer engagement. CRM strategy, based on practices and technologies, aims to manage customer relationships in a way that maximises customer value throughout their relationship with the company. However, according to the research findings, social media has taken over the management of CRM relationship from the managers who focused on the responses required to manage the customer (Baird & Parasnis, 2011).

Under the conventional CRM model, the company is in possession of extensive information about its customers, which it uses to manage relationships with those customers (Malthouse et al., 2013). Companies have adopted this strategy with the aim of maximising the lifetime value (CLV) of their customers, and thus their customer value, by using information about their customers. According to Malthouse et al. (Malthouse et al., 2013), in the conventional method, companies have the opportunity to invest more resources in certain customer groups, cross-sell to some groups, up-sell to others, and focus on reducing the cost of serving others. In this state, the company is the key player and talks to inactive customers. The ability of these customers to respond to the company's efforts is largely reflected in their buying behaviour. However, all of these things have evolved over the years with the development of modern technologies, such as online social networks. Customers and their highly influential virtual networks based on social media now define the conversation, which can greatly influence marketing, sales and service efforts due to its unprecedented immediacy and reach. Currently, customers are driving the conversation, and their highly influential virtual networks are based on social media (Baird & Parasnis, 2011). The customer in the virtual world (social media worlds) not only has more information about competing products available everywhere on mobile devices, but also has the ability to easily express and share their opinions to a large audience. This makes it increasingly difficult for companies to manage the messages customers receive about their products or services. This new way of contact has led to a concept that several scholars call Social Customer Relations Management (SCRM). SCRM has emerged as the method of choice for managing the evolving nature of customer interactions and relationships (Baird & Parasnis, 2011). As defined by Trainor et al. (Trainor et al., 2014), SCRM is "the integration of traditional customer-facing activities, including processes, systems, and technologies with emergent social media applications to engage customers in collaborative conversations and enhance customer relationships".

These bulletin board entries are still viewed by contact centre staff, who then forward them to the company staff responsible for customer service. This practice continues to this day (Geierhos, 2011). According to the findings of Trainor et al, managers in today's virtual world, just like marketing managers in the past (late 1990s to early 2000s) who were involved in the widespread adoption of customer relationship management (CRM) technologies, are busy integrating emerging technologies, such as social media applications, with existing systems and practices to develop new capabilities that foster stronger relationships with customers. Traditional customer relationship management (CRM) has been combined with modern technologies, particularly social media, to create a new idea of CRM. This new CRM concept involves a more collaborative and network approach to managing customer interactions (Trainor et al., 2014). With the emergence of SCRM comes the additional opportunity to take an expanded CRM perspective that recognises new opportunities enabled by the technological and social changes brought about by social media applications. This opportunity presents itself as an extension of the CRM perspective (Trainor et al., 2014). Organisations need to acknowledge this change and adopt a new strategy known as Strategic Customer Relationship Management (SCRM). This strategy recognises that the role of the business is no longer to manage customers, but rather to facilitate a collaborative experience and dialog that adds value to customers. A critical first step in developing a social CRM strategy is to understand the value consumers place on your products and services, especially when they are in the unique environment of a social platform (Baird & Parasnis, 2011).

To compile the necessary information for this study, the authors interviewed four of the six major telecommunications companies in Ghana as primary sources of information. MTN Ghana, Tigo Ghana Limited, Vodafone Ghana Limited, and Airtel Ghana are the four companies in this sector of the communications industry. The collection of netnographic data began with the use of information from two different social networks. These were the companies' official Facebook pages and their official Twitter accounts, both accessible through the links provided. This section provides a brief history of the companies, including information about their social network pages that were used as the basis for data collection. Below is some background information on each of the four companies that served as the source for data collection. We chose MTN Ghana, Airtel Ghana, Tigo Ghana, and Vodafone Ghana as our four telecommunications providers in Ghana.

- MTN Ghana, often referred to as Mobile Telephone Network (Ghana), is part of the MTN Group and operates as a subsidiary. MN is a South African-based multinational mobile telecommunications company founded in 1994 and formerly known as M-Cell. The company is headquartered in Johannesburg, South Africa. The majority of the company's customers are in African countries, although it also operates in countries in Europe and Asia. MTN the company's global divisions have a combined 300 million users (as of September 2015), making it the largest mobile network operator in the world. The main focus of this research is on the operations of MTN Ghana, the company's wholly owned subsidiary in Ghana.
- Airtel Ghana limited is a part of Airtel Africa, a subsidiary of the Indian telecommunications company Airtel, which operates in 17 countries in Africa. Airtel Ghana limited is a wholly owned subsidiary of Airtel Africa. The company provides 2G, 3G or 4G services depending on the country in which it operates its GSM network. The company operates GSM networks in all countries. Airtel Ghana has more than 2 million members in Ghana and offers services such as Airtel Money, Internet, Voice and Text Services, Business & Personal Solutions, all over a network that is 3.75 times faster than other networks.
- Millicon is an international telecommunications and media company, and Tigo Ghana Limited is a subsidiary of Millicon. Tigo Ghana Limited is one of the most successful telecommunications service providers in Ghana. Tigo is the brand name under which Millicon operates in fourteen countries in Africa and Latin America with a total of 63 million customers. Millicom (Ghana) Limited was the first mobile network in Ghana. It was established in April 1992 and began offering its services. The total number of subscribers of Tigo Ghana Limited is over four million, and the company today has a market share of about 17%. In addition to telecommunications services, Tigo also offers other services, such as Tigo Money, a mobile money transfer service in Ghana.
- Vodafone Group Plc is a global British company specialising in telecommunications, headquartered in London. As of March 2014, Vodafone had a total of 434 million customers, making it the second largest mobile telecommunications company in the world by both number of users and amount of revenue in 2013. Vodafone is a global company that owns and manages networks in 26 countries and has partner networks in over 50 other countries. In 2008, Vodafone began operations in Ghana after reaching an agreement to purchase a 70% stake in Ghana Telecom for a total of $900 million on July 3, 2008. On April 15, 2009, Ghana Telecom and its mobile subsidiary OneTouch renamed themselves Vodafone Ghana. This change took effect immediately. At the end of February 2016, Vodafone Ghana had approximately 7,859,486 customers, representing a market share of 21.95%.

For this study, data was collected from the two social media accounts of the four companies over a three-month period. In analysing the Facebook data, both companies examined posts and interactions from Facebook posts made between January and March 2016, which corresponds to the first quarter of 2016. The period that began on March 1 and ended on April 15 was used to collect data from Twitter, using tweets and retweets sent from the site. The following table provides a brief overview of the information collected from the pages. It shows the interactions between the frequency of posts (see Figure 6), the number of likes (see Figure 7), the number of comments (see Figure 8), and the number of shares (see Figure 9). The first set of summaries refers to the company MTN, the second to the company Airtel, the third to the company Tigo, and the fourth to the company Vodafone. In the first column, a letter S means that the interaction does not contain any media; a letter I means that the interaction contains an image; and a letter V means that the interaction contains video footage.

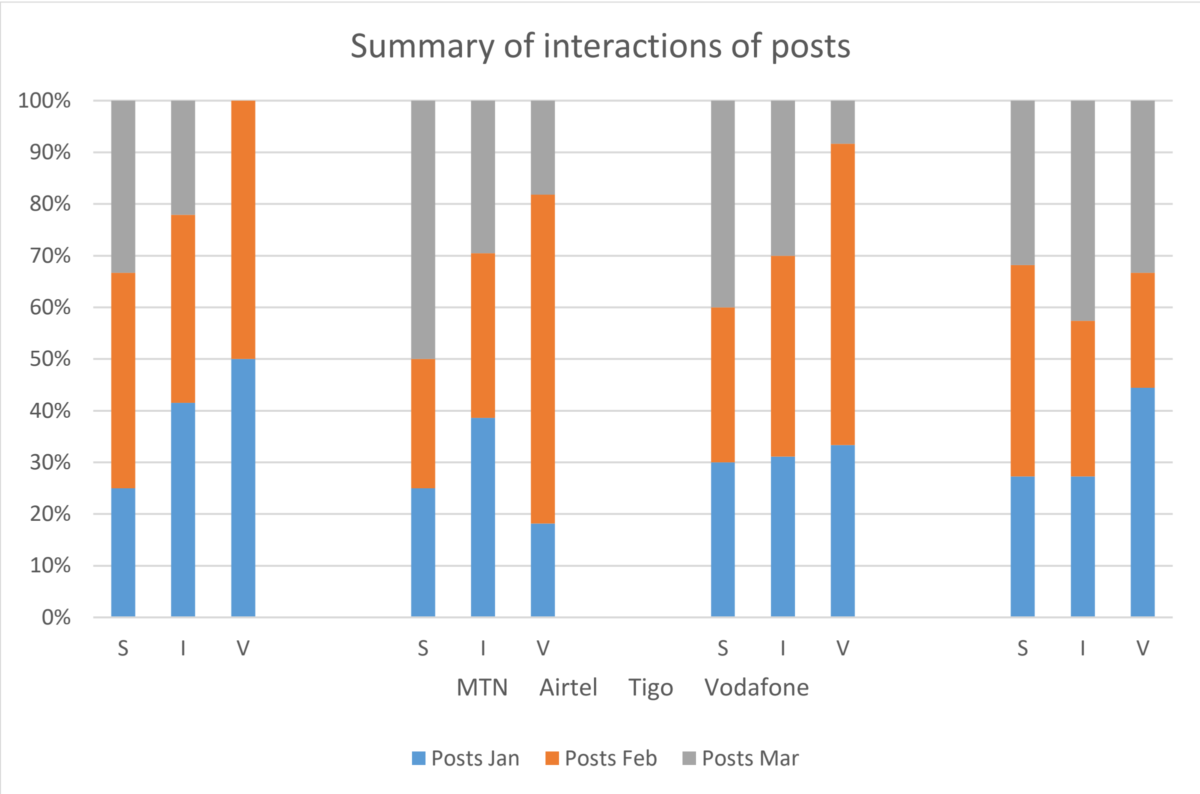


Figure 6: Summary of interactions of posts for all companies

MTN Ghana conversed online with its subscribers primarily through its Facebook page to disseminate information and respond to inquiries from customers. In these contacts, we reported network-related difficulties, shared updates on general concerns, and provided information on promotions and other programs of MTN Ghana. Most of these contacts took the form of status updates, either with photos, videos, or without any attachments. The interaction is now broken down into its component parts, as shown below. The vast majority of Airtel's postings consisted of status updates with pictures attached, which led to a significant increase in interactions among Airtel Network customers. The majority of customers' reactions to the postings were likes and comments.

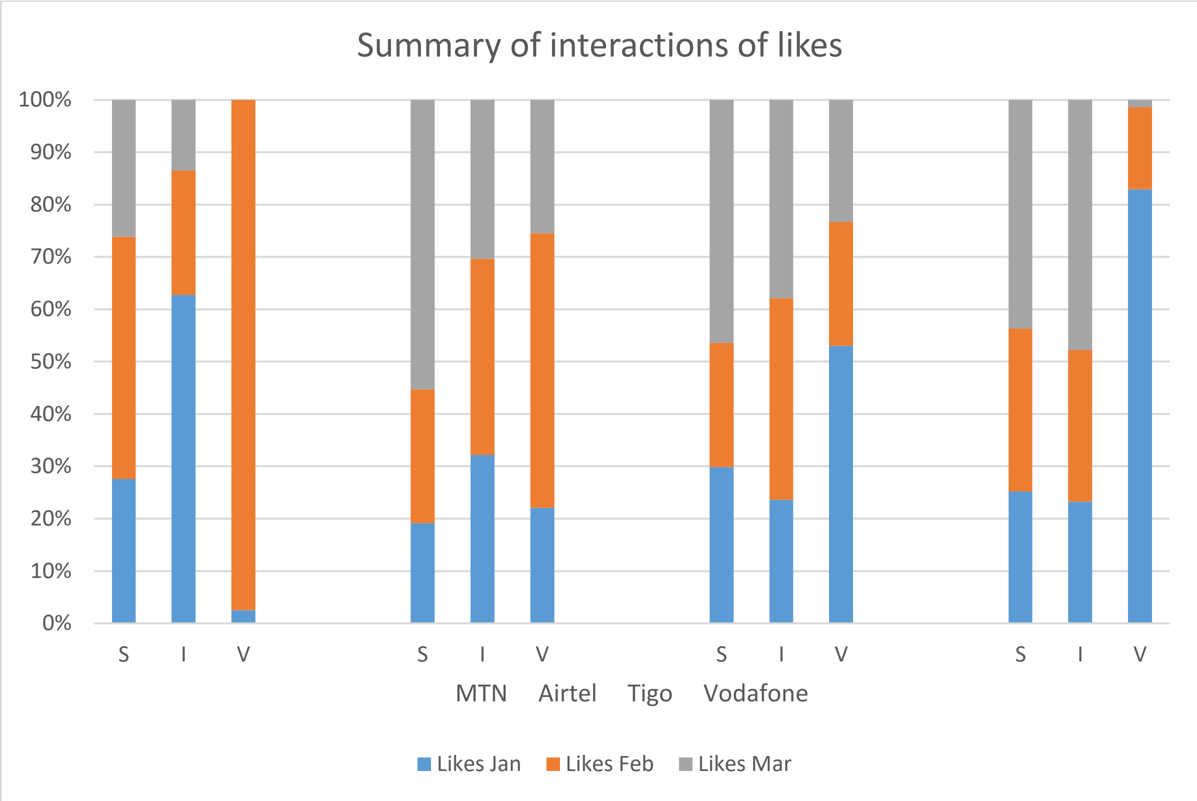


Figure 7: Summary of interactions of likes for all companies

Between January and March 2016, a total of 112 posts were uploaded to Tigo's Facebook page, including 12 videos, 90 photos, and 10 posts without media attachments. These posts received responses from subscribers in the form of 6665 likes, 3793 likes, and 4310 likes in January, February, and March, respectively; 1030 comments, 1343 comments, and 1119 comments in January, February, and March, respectively; and 2834 shares, 157 shares, and 286 shares in January, February, and March, respectively. Most of the comments and likes were generated by photo updates, while most of the video likes and likes were generated by image updates. In terms of the number of shares, video uploaded by the company was by far the most popular among subscribers.

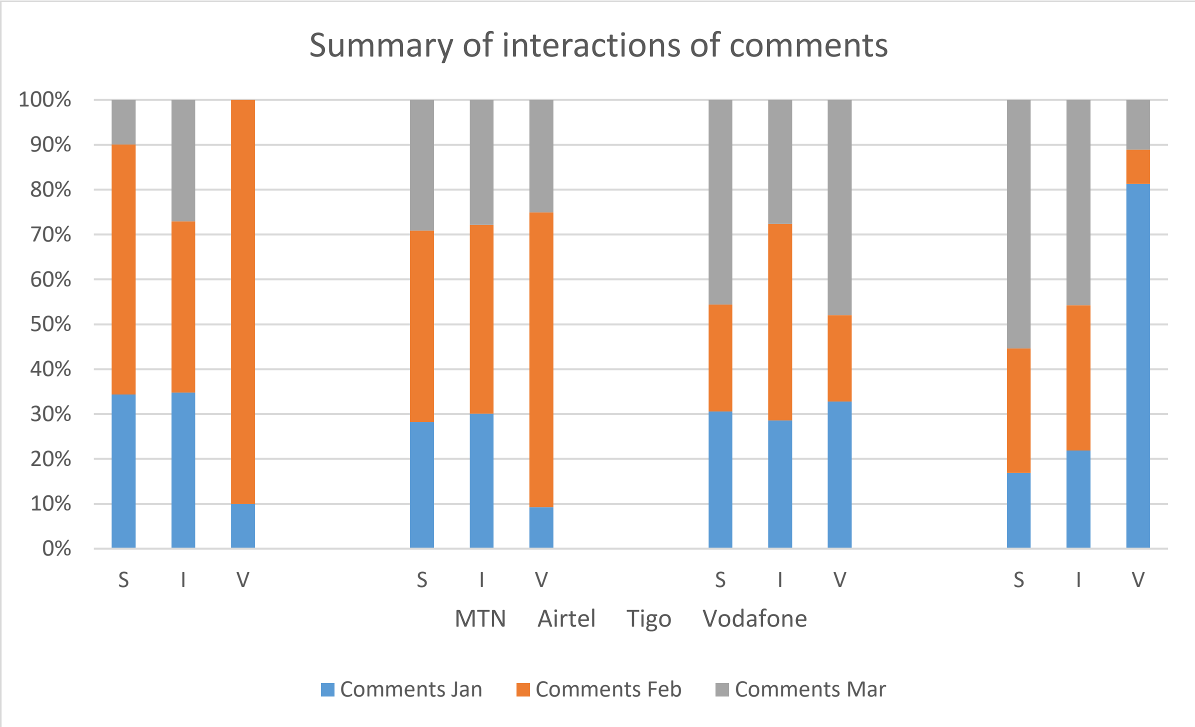


Figure 8: Summary of interactions of comments for all companies

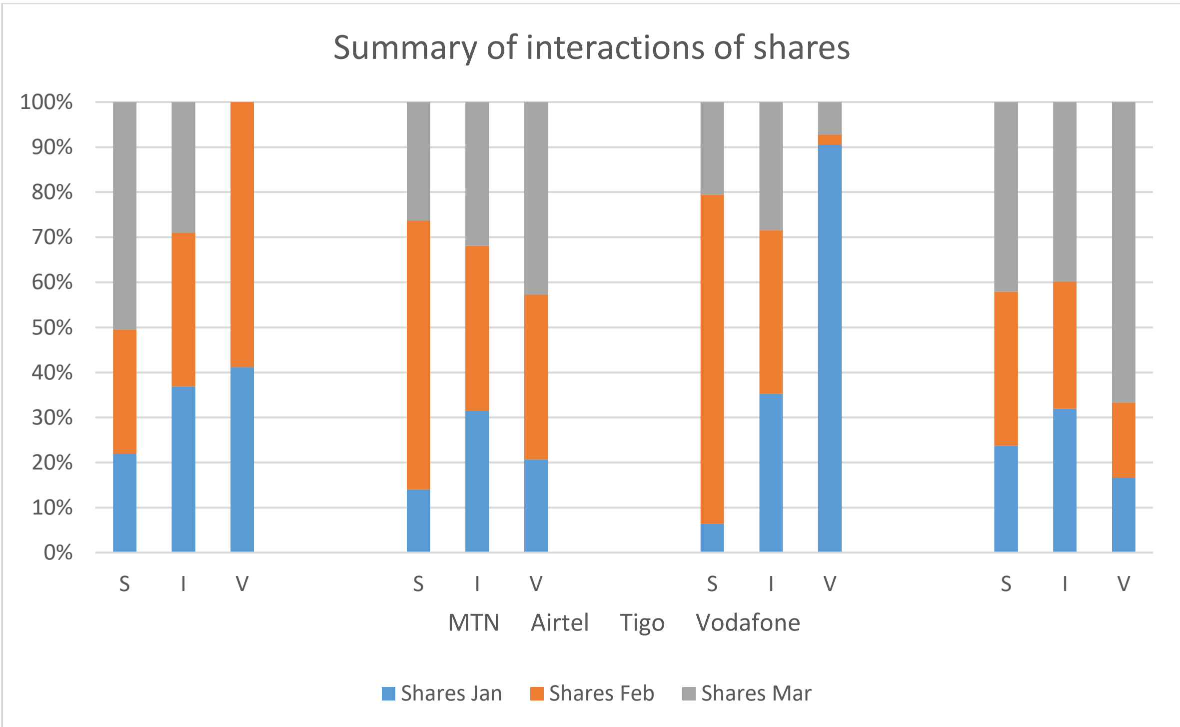


Figure 9: Summary of interactions of shares for all companies

The purpose of this study was to examine the extent to which telecommunications companies in Ghana, located in West Africa, use social media as a medium to interact with their customers.

In this chapter, we review the extent to which the objectives of the study were achieved in order to document the findings of the study and make suggestions based on them. In addition, we provide some suggestions for improving the use of social media as a tool for interacting with customers, as well as some suggestions for further research. As discussed in the previous chapter, Ghanaian telecommunications companies use social media to varying degrees as a medium for customer engagement. This use of social media is found throughout the country. The results show that some companies are very successful in engaging with their customers through social media, while others perform rather poorly and only engage with a small number of customers. While some companies are successful in this endeavour, others fail miserably. The research led to the following findings.

Social networks are beginning to be used as a customer interaction tool for telecommunications companies in Ghana. This is because of the realisation that they can serve as a means of instant and real-time interaction with customers. This is because of the realisation that social media can serve as a means of instant and real-time interaction with customers. Through social media, telecoms connect with 0.9% to 6% of their subscribers, while interactions in the form of inquiries occur on average with less than 2% of their customer base. MTN Ghana had the highest interaction with about 6% of its customers and received inquiries from 1.08%, while Airtel Ghana limited had the lowest form of customer interaction via social media with less than 1% of its subscribers. The four companies used for this research are: MTN Ghana, Airtel Ghana limited, and Vodafone Ghana. Although Airtel Ghana has a larger number of subscribers than Vodafone Ghana Limited and Tigo Ghana Limited, Airtel has the fewest interactions, suggesting that the company is not performing as well as it could be in terms of consumer interactions via social media.

Facebook has been by far the most popular platform for customer engagement, which is due to the fact that members can quickly see what interactions are taking place between other subscribers, which in turn leads to further interactions. Customers who see social media as a platform through which they can ask their questions use social media at a higher rate than customers who do not. Images, videos, and text messages are the three main forms of communication that can be generated through the use of social media with consumers. This process is initiated by telecommunications companies posting updates or status updates to which their customers can respond. Video updates, which often attract the greatest engagement from subscribers in the form of comments, tend to be the type of post that proves most successful. The study also concluded that despite Twitter being used as a platform for consumer engagement, the maximum number of characters that can be used on Twitter limits the overall usefulness of the network. According to the study's findings, companies that want to make the most of social media to engage with their customers need to ensure they have the appropriate teams in place to respond immediately to customer questions or concerns on social media. The following suggestions are based on the findings of this research: telecom companies need to find more ways to increase the amount of interaction between customers and companies on social media, and this can start with companies being innovative in their approach. Companies need to find more ways to increase the amount of interaction between customers and businesses on social media. Implementing such forward-thinking strategies requires companies to initiate consumer interactions by suggesting to customers or subscribers that they should use social media as their first point of contact for questions. This is because the solutions offered to subscribers can be easily viewed and used by other subscribers with similar problems. According to the findings of this study, companies that are not taking full advantage of social media should consider incorporating social media into their customer engagement tools. Not only is this a cost-effective strategy, but it also helps build more meaningful connections with subscribers. This study suggests that when developing marketing campaigns for their subscribers, companies should use the media, such as videos, that are most likely to capture those subscribers' attention. This will make them more likely to share things on social media with their other friends, ultimately leading to more interactions. The study also recommends that telecoms establish a specialised team that is always ready to respond to consumers' questions or concerns about social media in real time or as quickly as practical. This is one of the recommendations made by the study. Communications companies that want to make the most of social media can build a positive reputation by responding promptly to customer questions and concerns. Customers will be more willing to use social media as a tool to contact the company if they are assured that their questions will be answered (as quickly as possible) on such platforms.

Acknowledgment

This paper is partially supported by the EEA CZ-EDUCATION of the Project Activating Participants Through Social Networks and Gamification in Undertourism Areas with No: EHP-CZ-ICP-3-003

References

Alahakoon, T., Tripathi, R., Kourtellis, N., Simha, R., & Iamnitchi, A. (2011). K-path centrality: A new centrality measure in social networks. In *Proceedings of the 4th Workshop on Social Network Systems*, 1-6.

Albert, R., Jeong, H., & Barabási, A.-L. (1999). Diameter of the world-wide web. *Nature*, 401(6749), 130-131.

Baird, C. H., & Parasnis, G. (2011). From social media to social customer relationship management. *Strategy & Leadership*.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*, 22(6), 725-730.

Bindu, P. V., & Thilagam, P. S. (2016). Mining social networks for anomalies: Methods and challenges. *Journal of Network and Computer Applications*, *68*, 213-229.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology, 25*(2), 163-177.

Burt, R. S. (1992). Structural holes: Harvard university press.

Colquitt, J. A., & Zapata-Phelan, C. P. (2007). Trends in theory building and theory testing: A five-decade study of the Academy of Management Journal. *Academy of Management Journal, 50*(6), 1281-1303.

Dreyfus, H., Dreyfus, S. E., & Athanasiou, T. (2000). Mind over machine. *Simon and Schuster*.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68-77.

Eaton, J., & Kortum, S. (2002). Technology, geography, and trade. *Econometrica, 70*(5), 1741-1779.

Erdős, P., & Rényi, A. (1966). On the existence of a factor of degree one of a connected random graph. *Acta Mathematica Academiae Scientiarum Hungarica*, 17, 359-368.

Everett, M., & Borgatti, S. P. (2005). Ego network betweenness. *Social Networks*, 27(1), 31-38.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215-239.

Freeman, L. C. (1980). The gatekeeper, pair-dependency and structural centrality. *Quality and Quantity, 14*(4), 585-592.

Gallaugher, J., & Ransbotham, S. (2010). Social media and customer dialog management at Starbucks. *MIS Quarterly Executive, 9*(4), Article 3

Geierhos, M. (2011). Customer interaction 2.0: Adopting social media as customer service channel. *Journal of Advances in Information Technology, 2*(4), 222-233.

Geyer, C. J., & Thompson, E. A. (1992). Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society Series B-Methodological*, 54(3), 657-699.

Goh, K.-I., Oh, E., Kahng, B., & Kim, D. (2003). Betweenness centrality correlation in social networks. *Physical Review E*, 67(1), 017101.

Gould, R. V., & Fernandez, R. M. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology*, 89-126.

Gross, J. L., & Yellen, J. (2003). Handbook of graph theory. *CRC press*.

Gulati, R., & Westphal, J. D. (1999). Cooperative or controlling? The effects of CEO-board relations and the content of interlocks on the formation of joint ventures. *Administrative Science Quarterly*, 44(3), 473-506.

Helliwell, J. F. (2000). How much do national borders matter? *Brookings Institution Press*.

Hou, K., Hou, T., & Cai, L. (2021). Public attention about COVID-19 on social media: An investigation based on data mining and text analysis. *Personality and individual differences*, 175, 110701.

Hunter, D. R., & Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3), 565-583.

Jensen, D., & Neville, J. (2003). *Data mining in Social networks,* 287-302.

Jones, P. J., Ma, R., & McNally, R. J. (2021). Bridge centrality: A network approach to understanding comorbidity. *Multivariate Behavioral Research*, 56(2), 353-367.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59-68.

Kim, J. Y., Howard, M., Cox Pahnke, E., & Boeker, W. (2016). Understanding network formation in strategy research: Exponential random graph models. *Strategic Management Journal*, 37(1), 22-44.

Knoke, D., & Burt, R. S. (1983). Prominence. *Applied Network Analysis*, 195-222.

Kourtellis, N., Alahakoon, T., Simha, R., Iamnitchi, A., & Tripathi, R. (2013). Identifying high betweenness centrality nodes in large social networks. *Social Network Analysis and Mining, 3*(4), 899-914.

Kurtz, D. L., & Boone, L. E. (2012). Principles of contemporary marketing. *South-Western.*

Li, H., Wang, H., Yang, Z., & Odagaki, M. (2017). Variation autoencoder based network representation learning for classification. In *Proceedings of ACL 2017, Student Research Workshop*, 56-61.

Lopez-Castroman, J., Moulahi, B., Azé, J., Bringay, S., Deninotti, J., Guillaume, S., & Baca-Garcia, E. (2020). Mining social networks to improve suicide prevention: A scoping review. *Journal of Neuroscience Research*, 98(4), 616-625.

Malthouse, E. C., Haenlein, M., Skiera, B., Wege, E., & Zhang, M. (2013). Managing customer relationships in the social media era: Introducing the social CRM house. *Journal of Interactive Marketing*, 27(4), 270-280.

Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business horizons, 52*(4), 357-365.

McCallum, J. (1995). National borders matter: Canada-US regional trade patterns. *The American Economic Review*, 85(3), 615-623.
Molloy, M., & Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3), 161-180.
Moreno, J. L. (1934). Who shall survive?: A new approach to the problem of human interrelations. *Nervous and Mental Disease Publishing Co*.
Muniz, A. M., & O'guinn, T. C. (2001). Brand community. *Journal of Consumer Research*, 27(4), 412-432.
Naeem, S. B., Bhatti, R., & Khan, A. (2021). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*, 38(2), 143-149.
Nettleton, D. F. (2013). Data mining of social networks represented as graphs. *Computer Science Review*, 7, 1-34.
Nemes, L., & Kiss, A. (2021). Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication*, 5(1), 1-15.
Nieminen, J. (1974). On the centrality in a graph. *Scandinavian Journal of Psychology*, 15(1), 332-336.
Okamoto, K., Chen, W., & Li, X.-Y. (2008). Ranking of closeness centrality for large-scale social networks. In *Proceedings of the International Workshop on Frontiers in Algorithmics*, 186-195.
Pattison, P., & Robins, G. (2002). Neighborhood-based models for social networks. *Sociological Methodology*, 32, 301-337.
Rauch, J. E., & Casella, A. (2003). Overcoming informational barriers to international resource allocation: prices and ties. *The Economic Journal*, 113(484), 21-42.
Roberts, S. G., & Dunbar, R. I. (2011). Communication in social networks: Effects of kinship, network size, and emotional closeness. *Personal Relationships*, 18(3), 439-452.
Robins, G., Pattison, P., & Wang, P. (2009). Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. *Social Networks,* 31(2), 105-117.
Rosenkopf, L., Metiu, A., & George, V. P. (2001). From the bottom up? Technical committee activity and alliance formation. *Administrative Science Quarterly*, 46(4), 748-772.
Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4), 581-603.
Sashi, C. M. (2012). Customer engagement, buyer-seller relationships, and social media. *Management Decision*, 50(2), 253-272
Sharma, S., & Sharma, S. (2021). Analyzing the depression and suicidal tendencies of people affected by COVID-19's lockdown using sentiment analysis on social networking websites. *Journal of Statistics and Management Systems*, 24(1), 115-133.
Snijders, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2), 1-40.
Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409), 204-212.
Tan, Q., Liu, N., & Hu, X. (2019). Deep representation learning for social network analysis. *Frontiers in Big Data*, 2, 2, 1-10.
Trainor, K. J., Andzulis, J. M., Rapp, A., & Agnihotri, R. (2014). Social media technology usage and customer relationship performance: A capabilities-based examination of social CRM. *Journal of Business Research*, 67(6), 1201-1208.
Vollmer, C., & Precourt, G. (2008). Always on: Advertising, marketing, and media in an era of consumer control: *McGraw Hill Professional*.
Wasserman, S., & Robins, G. (2005). An introduction to random graphs, dependence graphs, and p*. *Models and Methods in Social Network Analysis*, 27, 148-161.

Wang, X., Lu, Y., Shi, C., Wang, R., Cui, P., & Mou, S. (2020). Dynamic heterogeneous information network embedding with meta-path based proximity. *IEEE Transactions on Knowledge and Data Engineering*, 34(3), 1117-1132.

Weidenbaum, M. L., & Hughes, S. (1996). The bamboo network: How expatriate Chinese entrepreneurs are creating a new economic superpower in Asia. *Simon and Schuster*.

West, D. B. (2001). Introduction to graph theory (Vol. 2). Prentice hall Upper Saddle River.

Wu, A., DiMicco, J. M., & Millen, D. R. (2010). Detecting professional versus personal closeness using an enterprise social network site. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1955-1964.

Xie, Y., Yu, B., Lv, S., Zhang, C., Wang, G., & Gong, M. (2021). A survey on heterogeneous network representation learning. *Pattern Recognition*, 116, 107936.

Yang, W. S., Dia, J. B., Cheng, H. C., & Lin, H. T. (2006). Mining social networks for targeted advertising. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, 1-10.

Ye, Q., Gao, Y., Zhang, Z., Chen, Y., Li, Y., Gao, M., ... & Chen, Y. (2022). Modeling Access Environment and Behavior Sequence for Financial Identity Theft Detection in E-Commerce Services. In Proceedings of the *International Joint Conference on Neural Networks*, 1-8.

Zhou, X., & Zafarani, R. (2018). Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, *2*.

Zhou, F., Liu, L., Zhang, K., Trajcevski, G., Wu, J., & Zhong, T. (2018). Deeplink: A deep learning approach for user identity linkage. In Proceedings of the *IEEE INFOCOM*, 1313-1321.

Zucco, C., Calabrese, B., Agapito, G., Guzzi, P. H., & Cannataro, M. (2020). Sentiment analysis for mining texts and social networks data: Methods and tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(1), e1333.

AI-Empowered Decomposition and Evolutionary Computation for the Closed Pattern Discovery

Jerry Chun-Wei Lin[a]

[a]*Department of Computer Science, Electrical Engineering and Mathematical Sciences Western Norway University of Applied Sciences, Bergen, Norway*

Abstract

In recent years, HUIM has become an important topic, especially in the field of basket market analysis. This is due to the fact that HUIM detects information or products that are useful for decision making, which is why it has become such a prominent topic. The process of extracting high utility itemsets from datasets has been the subject of several studies, and as a result of these studies, a considerable amount of sample information has been found. However, this technique is not able to provide correct options in a short time, such as real-time and online decision systems, because it is impossible to extract relevant and important information from a large body of knowledge in a short time. This is because both real-time and online decision systems operate in real time, which explains why this is the case. Closed high-utility itemset mining, is an approach to market engineering that discovers valuable patterns while reducing the total number of patterns discovered. However, previous research has shown that this approach is unable to manage enormous amounts of data, making it useless for today's Internet of Things (IoT) environment, where enormous amounts of data are collected every second. In today's world, there are billions of networked devices, each generating its own unique data. To begin, we will discuss the multi-objective model we have developed for mining closed data sets with high utility. The developed model takes advantage of Spark's MapReduce framework, which can speed up the mining performance. In addition, artificial intelligence has transformed the industrial operations, and one of the important applications of artificial intelligence is reducing the computational costs of optimization. To better and fast discover the required patterns, we apply the multi-objective k-means model to classify the transactions according to the significant connection that these transactions have with the frequency component. The MapReduce model and the genetic algorithm (GA) are used to explore potential candidates for mining closed high-utility itemsets in a large database. The results of a series of experiments have shown that the proposed framework performs better than the traditional models in terms of runtime, memory consumption, and scalability.

Keywords: artificial intelligence, optimization, decomposition, clustering, evolutionary computation.

1. Introduction

Artificial intelligence has transformed the industrial operations, and one of the important applications of artificial intelligence is reducing the computational costs of optimization. One of the best known optimization algorithms, also known as the Genetic Algorithm and abbreviated as GA, is said to have been developed by John Holland [14]. Finding optimal and feasible solutions that can be applied to NP-hard problems is a simple step in the process of building a model based on GA. Research projects using GA models have been conducted in a variety of fields and applications, each requiring a number of factors and techniques to be considered. In other words, working with GA is fast and straightforward. Moreover, despite the fact that it requires more time for convergence, it is quick to implement and can be adapted to a variety of circumstances, applications, and areas of study.

As a direct result of the rapid development of information technology, techniques fundamental to knowledge discovery in databases (KDD), such as association rule mining (ARM) or frequent item set mining (FIM), are being used in a wide variety of domains and applications [26]. The first method to discover association rules (ARs) in databases in a level-wise manner was called Apriori. This approach, which requires additional processing to create and evaluate candidates at each level, was the first technique to do so. It was also the method that required the most work by repeating database scans and candidates evaluation. It was proposed to use FP-growth [15] to get around this limitation and build a more compact tree structure. This structure would always track 1-itemsets and continuously determine frequent patterns from the FP-tree when implemented. The downward closure (DC) property has recently been extended to different models for a variety of information discovery tasks. Some examples include sequential pattern mining [2] and high-utility itemset mining (HUIM) [18, 20]. The problem of combinatorial explosion in a large search space is then approached from different aspects using a wide range of pattern mining strategies. Moreover, different pattern mining strategies are developed to solve the problem of combinatorial explosion in a large search space.

HUIM considers both the total amount and the profit made per unit when determining the most profitable database records. This helps ensure that the right decisions are made. On the other hand, this tactic can be used for a broader range of events and circumstances, provided a number of additional properties and constraints are considered. Producing High Transaction-Weighted Utilization Itemsets (HTWUIs) that exhibit the Downward Closure property required the development of a two-stage approach. This methodology uses transaction utility, abbreviated as the *TU*. This was done to narrow down the broad area searched [18], which is referred to transaction-weighted downward closure (TWDC) property. The results of the above studies have directly led to the in-depth investigation of a number of techniques to increase mining efficiency.

Some examples of these techniques are the HUP-tree algorithm [19], the UP-growth+ methodology [32], and the d2HUP algorithm [21]. In addition, the utility idea has been applied to a variety of domain-specific activities, resulting in improvements in the effectiveness of the patterns found [24, 25, 30].

Although HUIM can provide additional information for effective decision making in a particular area, such as market engineering, it is time consuming to analyze the vast amounts of patterns discovered by HUIM algorithms. This is especially important when a decision or business strategy needs to be made in a very short time. It was initially inspired by closed patterns, which retain a smaller number of decision principles, but are more concise. The process, known as closed high-utility itemset mining (CHUIM), involves searching databases to find closed high-utility itemsets (CHUIs) [33, 35]. Closed high-utility itemset mining, abbreviated CHUIM, is an approach that searches databases to discover closed high-utility itemsets, abbreviated CHUIs. As mentioned earlier, an itemset must satisfy the following two requirements to be called a CHUI: (1) it must not contain supersets that have the same support value, and (2) its utility must be greater than a specified utility threshold. The CHUI collection selected as a direct result will likely provide less information to decision makers, but it will be presented in a more streamlined form. The first approach used to obtain CHUIs was referred to as CHUD [33] and used a two-stage process [18] to extract the necessary data. This was necessary because the first stage generated multiple candidates and the second stage involved searching for the closed patterns, a process that takes a lot of time. This was necessary because this was necessary because this was necessary because this was necessary. To efficiently mine CHUIs by using the divide-and-conquer strategy, the CHUI-Miner model [35] was developed to use the extended utility (EU) list structure as a means to store sets of objects that are the result of transactions. This was to ensure that CHUIs could be mined in an effective manner.

To reduce the size of the area to be searched, the CLS-Miner [7] was developed. This model uses the matrix structure and is the most advanced mining technique for CHUIs currently available. The creation of the CLS-Miner enabled the successful completion of this task. On the other hand, this paradigm is unable to successfully manage large datasets due to its inherent limitations. By mining a collection of HUIs using an Apriori strategy [23] or a sampling model, the traditional HUIM demonstrated both distributed and parallel mining techniques [4]. The latter technique, which uses iterative MapReduce and has lower processing costs, is similar to the techniques used in mining frequent itemsets. The second technique, which parallelizes the HUI-Miner approach, uses a sampling model to estimate the estimated number of HUIs; as a result, fewer models are available for use in large datasets. This is because sampling models are less accurate than full models. In addition, a trustworthy model that can accurately distribute transactions among processing nodes is needed to finish mining the required CHUIs in an architecture that uses both parallelism and distributed computing. This is necessary to achieve the desired level of precision. In this paper, we provide a multi-objective k-means model for mining closed high-utility itemsets using the MapReduce paradigm. This model was developed to solve the identified problem. In the next part of this paper, we review the key findings that were obtained while conducting this research. Below is the contribution of these findings:

- We are developing a MapReduce architecture based on k-means with the goal of finding CHUIs present in large datasets.
- The model uses the multi-objective k-means algorithm to cluster the relevant transactions. This ensures that the total number of CHUIs detected is accurate and complete.
- The use of a model based on GA within the MapReduce framework enables rapid exploration of probable and likely patterns. This in turn leads to a significant reduction in the processing time required.
- The researchers found that the newly developed model was superior to the current state-of-the-art CLS-Miner in terms of runtime, memory utilization, and scalability across the number of nodes, which is the final analytical results from the experiments.

Section 2 then identified works relevant to the the current research. The past studies of the CHUIM were covered in Section 3. The content of the proposed model in the MapReduce framework was discussed in Section 4. The experimental results are discussed in Section 5. The conclusion is presented in the Section 6, followed by suggestions for further research is shown in Section 7.

2. Literature Review

This section contains a summary of research on evolutionary algorithms, high-utility itemset mining, and the MapReduce architecture.

2.1. Genetic Algorithm

One of the best known evolutionary algorithms, also known as the Genetic Algorithm and abbreviated as GA, is said to have been developed by John Holland [14]. Finding optimal and feasible solutions that can be applied to NP-hard problems is a simple step in the process of building a model based on GA. When we work with GA, each chromosome in a population serves as a solution, also for one of the individuals that are part of that population. In this step of the coding process, a search space must be coded for each of its possible responses. When using the GA, there are three different operations that are performed during the evolutionary development process. The genetic algorithm includes and uses these processes; these are crossover, mutation and selection. Then, the fitness function provided is used to evaluate the quality of the chromosomes. When the evolution process reaches a certain threshold for the number of generations or the fitness value converges, it is possible to stop the process and declare it complete. This can happen when the fitness value converges or when the number of generations reaches the threshold. Both events are possible [12]. Research projects using GA models have been conducted in a variety of fields and applications, each requiring a number of factors and techniques to be considered. In other words, working with GA is fast and straightforward. Moreover, despite the fact that it requires more time for convergence, it is quick to implement and can be adapted to a variety of circumstances, applications, and areas of study.

2.2. High-Utility Itemset Mining

Customers who store in the physical world often provide information about the value of products as well as the profit per unit for a number of different things through their purchasing behavior. This information can be used to make business decisions. In the past, traditional ARM or FIM research has focused solely on the frequency of item records; however, this does not provide sufficient information for decision making. Because of the high frequency with which frequent itemsets occur, the likely profit values of these sets are likely to be of lower magnitude. The high-utility itemsets was found using varied techniques in high-utility itemset mining (HUIM) [36]. We considered both the total purchase quantity and the unit profit of the items. There is a possibility that the traditional ARM and FIM methods of decision making will be replaced by HUIM since it provides more useful information for the final decision making.

A problem called a combinatorial explosion problem may occur. This is because HUIM does not contain the downward closure property [36]. Unsatisfied candidates were screened out using a two-step process known as TWU [18] that retained both the high transaction-weighted and the newly formed TWDC features to exaggerate the utility of an itemset. This approach was used to exclude candidates that did not meet the requirements. In this way, the overall utility of the itemset was increased. Because this feature provides higher utility value to the itemset, the search space that must be combed through to identify and remove the relevant HUIs is large. To this point, a significant number of HUIM models have been presented to the public and discussed. To search and find HUIs, researchers examined HUP-tree structure [19], utility pattern (UP)-growth, and UP-growth+ [32]. These techniques were used by the researchers to successfully locate HUIs. These algorithms continue to use the TWU model to discover a collection of HUIs, and then search the database to determine whether or not the candidates under investigation are valuable. The utility list structure for HUIM was created using HUI-Miner [20], which is more compressed and condensed to reduce the size for finding likely HUIs. This was done in place of the design and test technique. In the construction phase, a simple join operation is used to assemble the utility list structure to create k-itemsets. As part of the research conducted by FHM [8] , a matrix structure was developed to capture the occurrence of itemsets in conjunction with each other. This allowed investigators to immediately exclude non-desired alternatives from the search domain. It was decided to make accessible the most advanced generic HUIM strategy currently available, named EFIM [40]. This was done with the intention of effectively narrowing the search area to find satisfied HUIs.

Thus, if the threshold is not determined correctly, HUIM will create a larger number of HUIs, increasing the total number of HUIs. It can be difficult to perform analysis when trying to make the decisions based on a large number of samples, especially for online applications such as market analysis and forecasting. This can be very difficult to execute. By using a technique known as closed frequent itemset mining, which is both more effective and efficient, it is possible to create short and effective patterns for decision making. This is in contrast to the common practice of using entire patterns [16]. To make the patterns found in HUIM even more useful, a restriction to closed patterns was introduced in HUIM. This was done to increase the effectiveness of these patterns. A CHUI-Miner model [35], was created using a single-phase modeling strategy to make the process of finding closed high-utility itemsets (CHUIs) as fast and easy as possible. This is done by using the EU-list structure to keep track of upcoming shifts in the utility pattern. Rather than focusing on the candidate generation process, the challenge of discovering potential CHUIs is addressed using the divide-and-conquer technique. Since this part of the process depends on the TWU model to work properly, the mining phase requires the analysis of a significant amount of pattern data. In addition, the process of creating patterns in CHUI-Miner is likely to be a costly activity on your part. The utility list structure and EUCS have been paired to create CLS-Miner, which subsequently mines CHUIs in an effective manner. CLS-Miner, similar to CHUI-Miner, uses a single-phase model to find the necessary CHUIs. However, unlike CHUI-Miner, CLS-Miner uses its own pruning techniques to limit the number of possible patterns for CHUIs. The CLS-Miner is the most advanced and state-of-the-art CHUIM technology developed to date. On the other hand, the methods just discussed are only applicable to a single hardware and are not capable of managing huge datasets. These limitations were only one of several that were pointed out. The construction and implementation of a number of jobs has been completed, and efficient methods for information retrieval are now being created and improved [3, 11, 13, 27, 28, 37, 38].

2.3. MapReduce Framework

The term "MapReduce" refers to a method that Dean and Ghemawat developed in order to manage large datasets [5]. Two of the most important parts of this system are referred to as the mapper and the reducer. They make advantage of distributed and parallel models running on a cluster in order to achieve the goals they have set for themselves. If they are presented to it, the MapReduce architecture will treat the key-value pairs as input data and use them accordingly. Because the mapper and reducer approach data in different ways, the MapReduce architecture does not need any form of communication between the nodes that comprise the cluster. This is because the mapper and reducer each handle data in their own unique manner. As a consequence of this, the framework may now be distributed and carried out concurrently with a great deal less trouble. Overall, the reliability, flexibility, and parallelism of MapReduce make it a powerful programming framework that can be readily used in a broad range of different applications. MapReduce was developed by Google, which is capable of dealing with problems of a significant size in many domains and applications.

Lin et al. [22] developed SPC, FPC, and DPC with the intention of facilitating pattern mining within the MapReduce architecture. Later, Li et al. [17] developed the PFP approach, which parallelizes the FP-growth process but does not require remote computers to generate candidates. This allowed them to save time and resources. To extract HUIs from huge datasets, the Apache Hadoop framework PHUI-Growth, originally based on Apriori, was later enhanced and extended [17]. Since this technique requires significant additional computational effort, it is not effective on extremely large datasets. Chen et al. developed the HUI-Miner [4] approach by using Apache Spark as the primary tool. In this approach, sampling methods are used to summarize the provided data and discover HUI sentences. A sampling model has the potential to contribute significantly to performance improvement as it can provide an estimate of the total number of HUIs. However, this technique has not been able to provide accurate results in terms of the total number of HUIs or simply the utility of the entire collection of item sets. For this reason, the ability to make an accurate and precise judgment based on the patterns found and recognized is limited. This judgment would be based on the collected and recognized patterns.

Wu et al. [34] work on the fuzzy version of HUIM in distributed and parallel Hadoop systems. They concluded that HUIM and FIM can best be distinguished by doing their work in the domain of fuzzy set theory. The reason for this is that HUIM and FIM are often considered equivalent. They then developed a MapReduce framework based on fuzzy-set theory that is capable of discovering fuzzy utility itemsets from huge datasets in an efficient manner. To overcome the computational limitations that can be caused by very large datasets, a Hadoop-based version is now being investigated as a possible solution. The experimental study presented made it very clear that the fuzzy version of HUIM used in their study performed remarkably well in an artificially controlled environment. This was illustrated by the results of the study. However, it was still uncertain whether a method based on a single computer would work successfully with distributed infrastructures and the data itself. This was because there was no way to test this hypothesis. With the ever-growing CHUIM research, it is of paramount importance to develop a system capable of immediately identifying the collection of CHUIs in a large database. The reason for this is that the research is becoming more and more extensive.

3. Preliminary and Problem Statement

For illustration, consider the following collection of elements to keep in mind: I={i_1, i_2, …, i_m} in the database there is a positive value called $p(i_j)$, which represents the external or profit value. This value can be found after any of the previous four values. In other words, we can say that D consists of n transactions, where n is the total number of transactions. Besides the transaction ID, the T_d has its very own identity represented by the abbreviation TID. This identifier is separate from the transaction ID. The fact that a transaction T_d contains items that all have a positive value is denoted by the notation $q(i_j, T_d)$, which reads as follows: this value represents the monetary value of each item i_j

that was involved in the transaction T_d. Depending on the surrounding circumstances, this amount can be considered as the internal utility for the transaction denoted by T_d. For example, an itemset X is understood as a collection of k independent items, such as $X = \{i_1, i_2, \ldots, i_k\}$ and $X \subseteq I$. The length of X is referred to k, and X may also be referred to as k-itemset in some cases. The following table contains examples of everything that can be included and used in CHUIM.

Table 1: A quantitative dataset

TID	**Transaction**
T_1	(a,3), (c,3)
T_2	(b,4), (c,4), (e,3), (g,1)
T_3	(a,2), (b,3), (e,2), (f,2), (g,3)
T_4	(a,3), (c,2), (d,4), (e,3)
T_5	(c,1), (e,3), (g,2)

Table 2: The unit of profit table

Item	a	b	c	d	e	f	g
Unit of Profit	2	3	4	2	1	5	3

Table 1 is a transactional database and has five distinct transactions with values T_1, T_2, T_3, T_4 and T_5. The database stores seven distinct objects, each of which has its own unique identifier: a, b, c, d, e, in addition to f and g. These identifiers are as follows: The information you are looking for about the unit profits of the goods is in Table 2. The following items are included in this information: a is set as 2, b is set as 3, c is set as 4, d is set as 2, e is set as 1, f is set as 5, and g is set as 3, respectively.

Definition 1. *The following explains what the notation $u(i_j, T_d)$ represents: the utility that an object i_j has when considered in the context of a transaction T_d as:*

$$u(i_j, T_d) = q(i_j, T_d) \times p(i_j), \tag{1}$$

in which $q(i_j, T_d)$ shows the internal utility (also called quantity) of $i_j \in T_d$, and $p(i_j)$ shows the external utility (also called profit unit) of i_j in the predefined profit table.

Definition 2. *The notation $u(X, T_d)$ expresses the utility of an itemset, focusing on X, in a transaction T_d according to the following definition:*

$$u(X, T_d) = \sum_{i_j \in X} u(i_j, T_d) \tag{2}$$

Definition 3. *In a database, the utility of an itemset is denoted by the symbol $u(X)$, which has the following meaning according to its definition:*

$$u(X) = \sum_{X \subseteq T_d \wedge T_d \in D} u(X, T_d) \tag{3}$$

Definition 4 (Transaction Utility). *Below is the definition of the notation used to express the utility of the transaction, written as $tu(T_d)$:*

$$tu(T_d) = \sum_{i_j \in T_d} u(i_j, T_d) \tag{4}$$

Definition 5 (Total Utility). *The following equation, given by the notation $u(D)$, can be used to express the total utility of a database:*

$$u(D) = \sum_{T_d \in D} tu(T_d) \tag{5}$$

Definition 6 (High utility itemset mining). *According to the formula we will present in a moment, an itemset X is considered as a qualified HUI if the following condition is met as:*

$$u(X) \geq \delta \times u(D) \tag{6}$$

Definition 7 (Closed high utility itemset, CHUI). *If another superset Y does not have the same support as the element set X, then the HUI status of the element set X is considered closed (CHUI). Never forget that Y can also be a HUI.*

In order to filter down the pool of candidates that do not hold promise, both the classic ARM and the FIM are used, both of which maintain the DC property. This is done in order to narrow down the search space. The issue of "combinatorial explosion" in the effective reduction of HUIs to preserve the DC property was addressed by the introduction of the TWDC property as well as the two-phase method. The idea of two-phase model to keep the DC property of the itemsets is then described below.

Definition 8. *Given by the notation $twu(X)$, which is used to express the twu value of an itemset called X as:*

$$twu(X) = \sum_{T_d \in D \wedge X \subseteq T_d} tu(T_d) \tag{7}$$

When it comes to mining HUIs, it is important to remember that the TWU model is used by both CHUIM and HUIM [8, 18, 32]. It is also possible to identify the collection of HUIs by using the well-known utility list structure used in HUI-Miner and the EUCS structure used in FHM [8]. In the utility-list structure, the join approach is used to construct the promising itemsets one level at a time. This is done to maximize efficiency. This structure contains the information needed for iterative progress and it is very important. Since EUCS is a matrix that stores the *twu* values of 2-itemsets, it is possible to find unsatisfied 2-itemsets early in the mining process. This is made possible by the fact that EUCS stores these values. This can reduce the number of candidates for subsequent generations when the number of 2-itemsets is reached.

CHUI is not only useful for online assessment, but also provides a small amount of information about HUIM that is complete despite its modest size. This information can be found here. Currently, there are no studies that evaluate the potential of CHUIM to handle huge datasets. This is because most recent research has looked at generic and conventional models [7, 35]. This paper aims to develop a MapReduce architecture capable of processing CHUIM in a large database in an effective and efficient manner.

Problem Statement: Suppose you have a database, which you call D, and you also have a profit table, which you call *ptable*. The threshold for the smallest utility is given by the symbol *delta*. In the context of this study, the CHUIM problem is to obtain an accurate and complete collection of CHUIs from a huge database based on the cloud computing models.

4. Developed CHUI Mining Algorithm

Figure 1 shows both a GA-based model and a 3-tier MapReduce architecture for processing large amounts of data. It uses the multi-objective k-means model to provide better results.

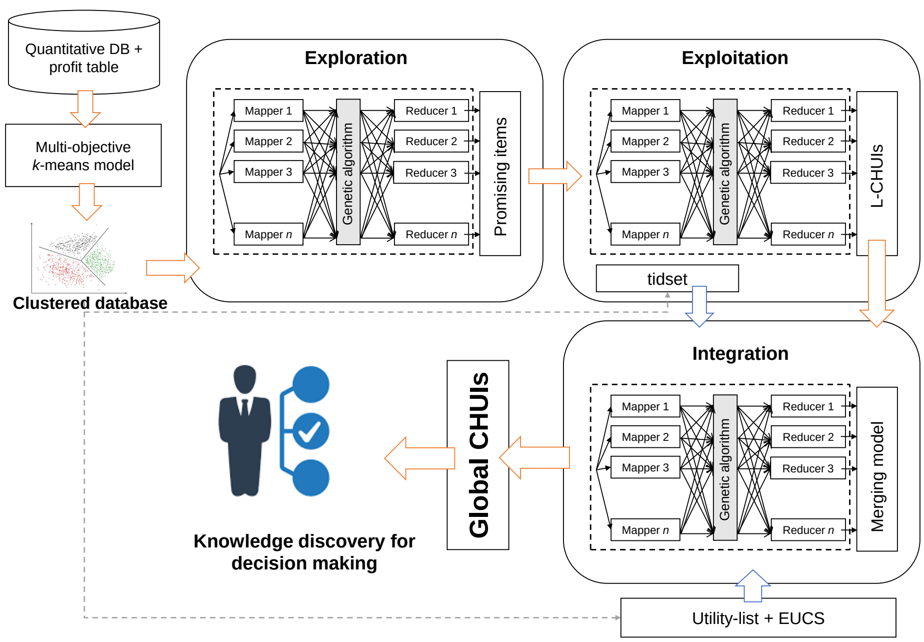


Figure 1: The flowchart of the designed approach.

4.1. Decomposition

The transactions in D are used to form the groups represented by the notation $G = \{G_1, G_2, \ldots, G_k\}$, where each group denoted by the notation G_i

is a subset of the transactions and k is the total number of groups created. Since none of the groups are related in any way, the following rule applies to all possible permutations:

$$(G_i, G_j), I(G_i) \cap I(G_j) = \emptyset, \tag{8}$$

in which $I(G_i)$ shows the set of items in a group G_i.

Proposition 1. *It is subdivided into the many types of transactions that can be performed with D. If the groups in G have no items in common, then the only way to determine the relevant frequent itemsets is to take the union of all the relevant frequent item sets for the groups. As a direct result, it is possible that the following will occur:*

$$F = \{\bigcup_{i=1}^{k} F_i\}, \tag{9}$$

in which there is a collection of frequent itemsets for the group G_i, which is called F_i.

Proof 1. *Consider $\forall(i,j) \in [1...k]^2$ $I(G_i) \cap I(G_j) = \emptyset$, we then can have that $\forall i \in [1...k]$: $F_i = \{p | sup(D, I, p) \geq minsup\}$. The support of a pattern p is determined by checking all transactions in D. Considering a pattern p appears in $I(G_i)$, i.e., $p \subseteq I(G_i) \Rightarrow \forall e \in p, e \in I(G_i) \Rightarrow \forall e \in p, e \notin I(G_j), (\forall j \in [1...k], \vee j \neq i) \Rightarrow p \nsubseteq I(G_j) \Rightarrow F_i = \{p | sup(G_i, I(G_i), p) \geq minsup\} \Rightarrow F = \{\bigcup_{i=1}^{k} F_i\}$.*

To identify a fully dependent group, the transactions in D must satisfy all of the above characteristics. It is possible to find the significant frequency groups if the search is restricted to groups of transactions. Indeed, the purpose of this endeavor is to reduce the number of items that are common to all of the many transactions that have emerged as a result of this endeavor. In the past, k-means [10] and DBSCAN [29] have been shown to have suitable transaction decomposition performance, with k-means outperforming DBSCAN in terms of transaction decomposition performance. Thus, we select k-means as a main model used in the designed transaction decomposition framework. The goal of the framework is to group all high-value transactions into one group, and the k-means model is used for this purpose. The k-means is represented here in this section in several different iterations. The first approach considers not only the items that are shared across different groups of transactions, but also the items that are shared by members of the same group. In contrast, the second approach[31] uses both a genetic algorithm and the k-means concept simultaneously. The two different types of disassembly are broken down into their components as follows:

1. There are two different types of multi-objective k-means: One considers the number of common features between different groups and the other

considers the number of common variables between different transactions belonging to the same group. In both variants, the total number of common features and variables is taken into account. After the transactions have been classified into the appropriate groups, we can start the process. In most cases, we choose a sample that is typical of the transaction that is considered the geographic and conceptual core of each group. During each cycle of the iteration process, the transactions are each assigned to a group to which they are eventually distributed. This helps reduce the amount of things that are shared between different groups, while increasing the number of items that are shared within the transactions of a single group. It is necessary to perform this approach several times to ensure that the group continues to function properly.

2. Multi-objective GA-based k-means: Even though the concept of multi-objective k-means is relatively innovative, the process of decomposing complete transactions is very time consuming. This is especially important when dealing with a very large number of transactions. Using more sophisticated methods could allow faster completion of the computation process. In the next section, we would like to talk about the hybrid genetic methods we have been working on. This research focuses on the solution space of decomposition results, which considers all conceivable combinations of these results. Since there are k groups, each conceivable answer can be represented in terms of the k components that make up the solution. Each component represents a unique transaction that occurred within that group. The population of a genetic algorithm is initially inoculated with a predetermined number of different possible solutions. Over the course of the current generation, the overall population has not noticeably increased. Intelligent operators such as crossover and mutation are used in the search for a solution. We refer to the different types of groupings that share most characteristics as "crossover points" to facilitate a more comprehensive analysis. The group of transactions that share the fewest components is the one that is considered when calculating the exact time of mutation. The responses that have the potential to have the greatest impact are then passed from one generation to the next via the same genetic mechanism.

To mine closed patterns with a high utilization, a designed multi-objective evolutionary computation model makes use of multi-objective k-means, genetic algorithms (GA), and MapReduce. Algorithm 1 provides a representation of the pseudocode for the step known as the decomposition phase.

In the technique of Algorithm 1, there are two basic requirements, the first of which is to determine whether the chosen decomposition algorithm is the multi-objective k-means only or the combination with the genetic algorithm. Both alternatives can be considered. In the first alternative, we must first create the clusters and then distribute the transactions among the clusters in an iterative process, taking into account the multi-objective criteria. This process is repeated several times until the initially chosen target stopping point is

Algorithm 1: Decomposition step

```
Input: The set of transactions D.
Output: The set of clusters G.
if algorithm = "MultiObjectiveKmeans" then
    initialize_clusters(G, D);
    repeat
        for D_i ∈ D do
            assign(D_i, G);
            evaluate_multi_objective(G, D);
        check_stopping_criteria(G);
    until stopping_criteria(G, D) == true;
if algorithm = "MultiObjectiveGAKmeans" then
    initialize_population(S, D);
    initialize_iteration(iteration, 0);
    repeat
        crossover(S, D);
        mutation(S, D);
        fitness_multi_objective(S, D);
        selection(S, D);
        increment(iteration);
    until iteration = maximum_number_iterations;
    build_clusters(S, G);
return G;
```

reached. The stability of the clusters is the criterion used to decide whether it is appropriate to stop the process. In the second option, we must first initialize the population, then explore the solution space using the crossover and mutation operators, evaluate the generated solutions considering the multi-objective criteria, and finally select the most optimal solutions. The selection of the most optimal solutions is the final step. This process can be repeated an infinite number of times, as this is the maximum number of repetitions allowed. After that, we will start building the clusters by using the most efficient options. Following the breakdown, the designed algorithm is to find closed patterns with high utility using a three-stage MapReduce architecture, genetic algorithms, and a mixture of multi-objective k-means. The processes of exploration, exploitation, and integration are ultimately integrated into the developed framework. This is done to speed up the execution of mining activities [28].

4.2. Exploration

Each mapper in the chain is assigned a partition once the transactions are sorted into their respective categories and then further partitioned. At this stage, the MapReduce architecture is used to examine the components and the supersets of those components that may make up CHUI. This examination is performed to determine whether or not CHUI is present. The potentially productive groupings of elements and the supersets of these collections are not included in the mining process due to a discovery made during this stage of the process.

In short, the MapReduce algorithm first divides the clustered dataset into many different regions. This is done to facilitate the subsequent steps. After that, a new mapper is responsible for each of these regions. After this step, the model GA provides a list of reasonable possibilities that can be used to narrow down the search space. This will help speed up the process. In the subsequent MapReduce phase, all satisfied frequent records are discovered and all unpromising frequent records are removed from the database as needed. In the next iteration of the MapReduce for CHUIM, only closed frequent itemsets are allowed. Below is a breakdown of the exploitation phase into its individual elements, described in more detail below. More detailed information can also be found in [28].

4.3. Exploitation

In the phase referred to exploitation, the CHUIs for each partition are determined using the last iteration of the CHUIM models (e.g., CLS-Miner [7]). Using a second MapReduce, the fulfilled item sets from the first MapReduce are processed in parallel on each node that uses this second MapReduce. Since it is difficult to extract all CHUIs from the database, this is a necessary step. Since each node in the MapReduce architecture makes less use of RAM, this design makes it possible to run a very large database on a single server. On each node, the utility of each candidate is computed so that the progress of the exploitation mining process can be tracked and evaluated. This enables the tracking and evaluation of the process. In this particular situation, both the transaction ID and the frequently used itemsets are stored in a horizontal model called a `tidset`. The horizontal model is easy to offer an estimate of the amount of work that has already been done in mining.

In the second MapReduce task, a simple load balancing strategy is used to split the transaction data into more manageable chunks based on the size of each chunk. Then, these chunks are passed on to subsequent jobs. For this reason, the technique may require fewer computational resources to exploit the vulnerability. It is likely that the number of jobs that can be developed is in direct proportion to the number of people engaged in the mapping function. When processing a transaction, the weighting of each node is determined based on the number of promising item records. The most promising node is finally selected as the winner because it can distribute the computations fairly among all nodes. Compared to standard serialization practice, this approach can result in significant savings in terms of processing costs. The equation below illustrates this point very well.

$$WL_i = WL_i + Num, \tag{10}$$

where the value *Num* indicates the number of patterns created as a direct result of a single MapReduce operation, while the value WL_i reflects the total amount of work completed. Both values are denoted by the variable WL_i.

To determine the list of local CHUIs located within each partition, the CLS-Miner [7] is used. Each mapper receives the CHUI as a pair consisting of a (*pattern*, (*utility*,p_i)). This pair is mapped as a CHUI. For more information, see the sources given in [28].

4.4. Integration

The expansion of mining operations obscured some of the patterns in the earlier clusters, but in this new cluster they have all been revealed again. Examination of the common and clustered components of the resource occurs in both the exploration and exploitation phases. With this strategy, it should not be too difficult to detect the CHUIs that are relevant in the database. The production of CHUIs is possible through the use of common components. It is important to apply the integration function to determine whether or not the pattern can be applied to the entire transaction database. This can be done by determining whether it is possible to apply the pattern or not. It is recommended that the integration function aggregate across all clusters the total amount of local support for common patterns. This would cover all clusters. When creating the required patterns for the entire transactional database, the global CHUIs of the common elements will be linked to the local CHUIs of the clusters. This will ensure that the patterns created are accurate and comprehensive. In the second phase of the MapReduce algorithm, the `tidset` output is generated. This is done to reduce the amount of computer resources needed to mine the patterns of each node. Then, the data required for the computation is stored using these structures, which ultimately results in significantly reducing the overall cost of the computation. In [7], it contains more in-depth information about the `utility-list` and the `EUCS`.

After that, you can use the third MapReduce to find global patterns with respect to CHUIs by combining the `tidset` with the set of candidate patterns consisting of local CHUIs. Following this, a description of the mapper and reducer functions that make up the third iteration of the MapReduce framework is indicated in [28].

5. Experimental Evaluation

In the context of the three-tier MapReduce architecture, a comparison was made between the designed model and the CLS-Miner [9] in terms of runtime, memory consumption, and scalability over a variety of different node counts. Next, Spark employs the MapReduce paradigm, which, when applied to managing large databases, provides more flexibility and scalability. This is done by distributing the work across multiple nodes in a cluster. Next, these dataset properties are represented in Table 3, where $|D|$ represents the size of the database and indicates the total number of transactions contained in the database. The value in $|I|$ represents the total number of unique entries contained in the database. For transactions, the variable **C** indicated the normal number of entries, while the variable labeled **MaxLen** indicated the maximum storage capacity of the database. Then, the dataset referred to Table 3 is replicated several times in different steps (e.g., 1, 20, 50, 100, 200, 500, 2,000, 5,000, and 10,000 times) to evaluate the effectiveness of the databases.

Table 3: Characteristics of the conducted datasets

Dataset	$\vert D\vert$	$\vert I\vert$	**C**	**MaxLen**
BMS	59,601	497	2.5	267
SIGN	730	267	52.0	94
MSNBC	31,790	17	13.3	100
Leviathan	5,834	9,025	33.8	100

5.1. Quality of the Clustering

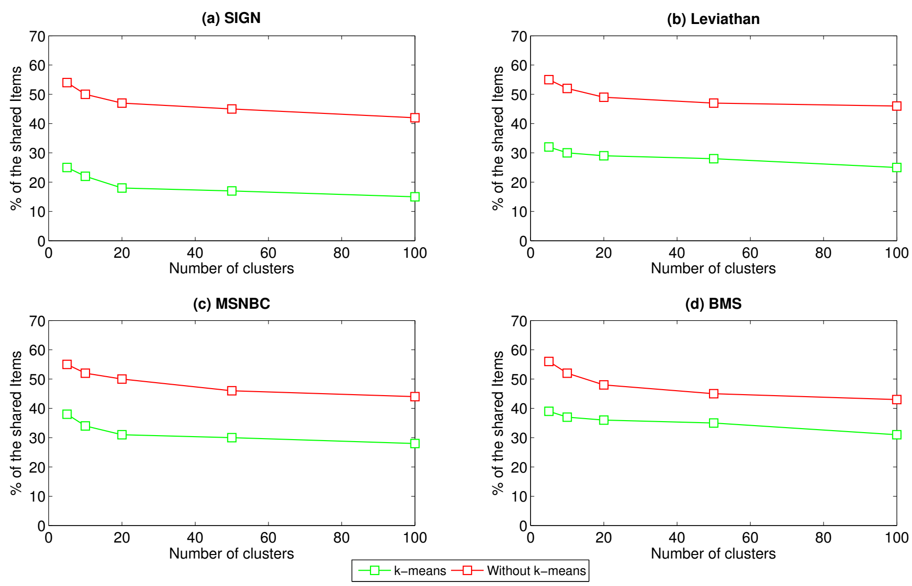


Figure 2: Decomposition clustering quality.

Figure 2 shows the results of the k-means clustering in addition to the results of the intuitive clustering for each of the four datasets used for the test. The transactions are clustered in a completely arbitrary manner using a technique known as intuitive clustering. If one wants to get the highest possible quality in the returned clusters, one must keep the percentage of identical components as low as possible. The percentage of overlapping features present in both methods decreases significantly when the items in question are divided into groups ranging in size from one to one hundred. There is a significant difference between

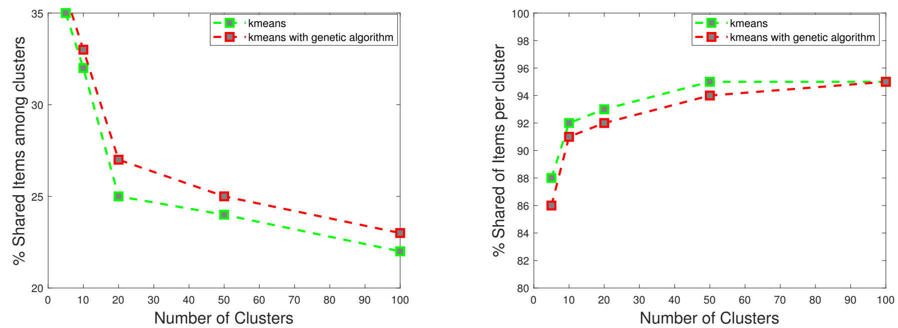


Figure 3: Decomposition quality in large-scale data (10 million of transactions).

the k-means algorithm and the intuitive strategy, even though both approaches have their positive sides. Unlike the intuitive technique, which uses only the random operator to complete the calculations, the k-means algorithm relies on the centroid and the similarity measure to reach its conclusions. A convincing example of the data decomposition approach known as was presented in the study. A second experiment was conducted with the goal of putting the decomposition strategies discussed in this study to the test. This experiment was conducted in accordance with the above objective. Both the number of clusters and the number of transactions were manipulated during this experiment. Using the IBM generator[1], we were able to generate 10 million unique transactions. The k-means-genetic approach produces a solution faster than the k-means algorithm when the number of clusters is between 5 and 100. This is true whether the number of clusters is increased or decreased (number of common elements between clusters and number of common elements within the transactions of a cluster). The genetic operators discovered during this study are absolutely necessary to reach this particular conclusion. When analyzing 10 million transactions, the genetic k-means algorithm is three times faster than the k-means approach.

5.2. Speedup and Memory Usage

A comparison is made between the created framework and the most advanced mining framework currently available, which is called CLS-Miner [7]. The comparison is made in terms of the speedup and memory consumption of the developed framework, both of which are shown in the Figures 4 and 5 to highlight the practicality of the developed framework. The comparison is in terms of the acceleration efficiency of the developed framework.

The speedup of these two algorithms was proportional to the number of nodes, which can range from one to thirty-two. The developed model is much more advanced than the CLS-Miner in every conceivable way. Responsible for

[1]https://github.com/zakimjz/IBMGenerator

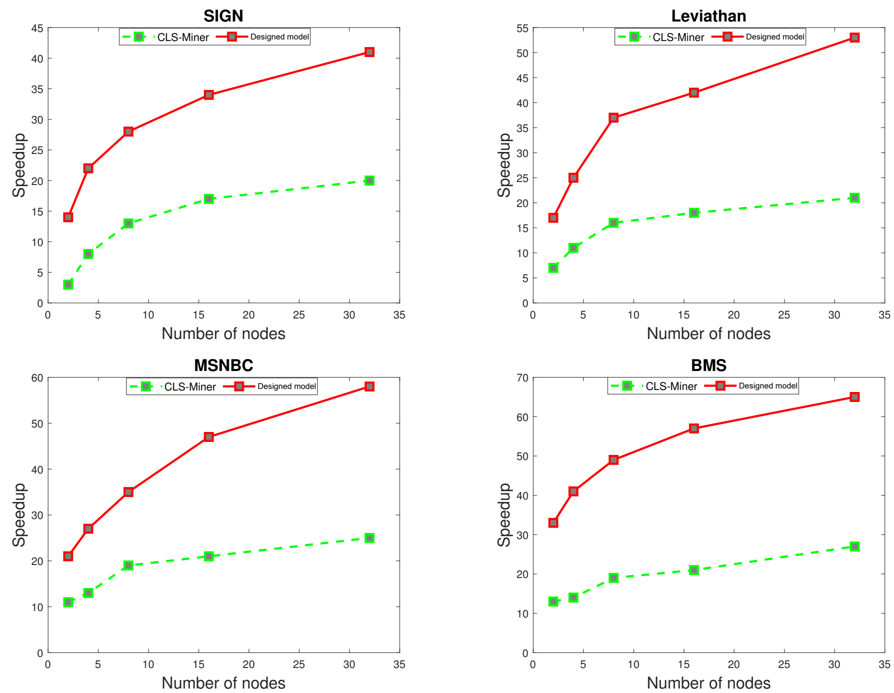


Figure 4: Speedup performance of the compared algorithms.

this are the decomposition, where each cluster contains similar transactions, and the clever operators of the evolutionary algorithm, which explores the accessible solution space in an appropriate way in a given time. The evolutionary algorithm also explores the accessible solution space in an appropriate manner in a given amount of time.

Next, the memory requirements of the two alternative methods must be analyzed and compared. In order to perform an in-depth analysis, the size of the datasets used was scaled to 10,000 times their original size. As can be seen in the Figure 5, the memory requirements of the CLS-Miner are significantly higher than those of the proposed model. This is the result of a type of algorithm known as a genetic algorithm, which analyzes only the solutions that are most relevant to the question at hand. The CLS-Miner requires a significant amount of memory to store the CHUI candidates. This is necessary because the CLS-Miner analyzes each individual transaction.

5.3. Scalability

To emphasize the usefulness of the developed technique, the scalability of the approach is illustrated using a large data set in Figure 6. The result is intended to illustrate the usefulness of the developed method. We planned to run a total of 1,000 tests to determine the scalability of our BMS dataset, with

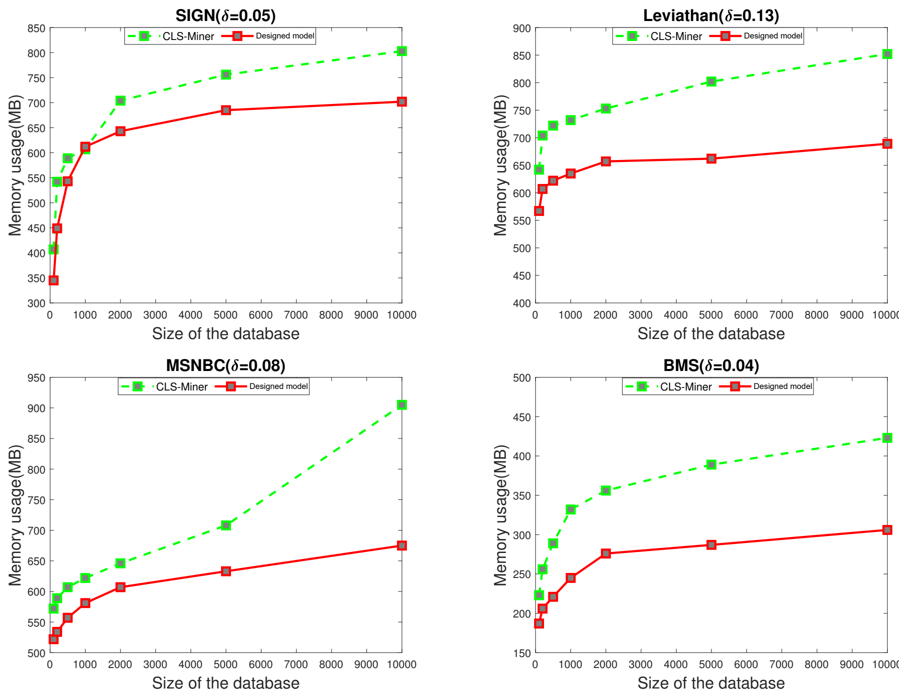


Figure 5: Memory usage of the compared algorithms.

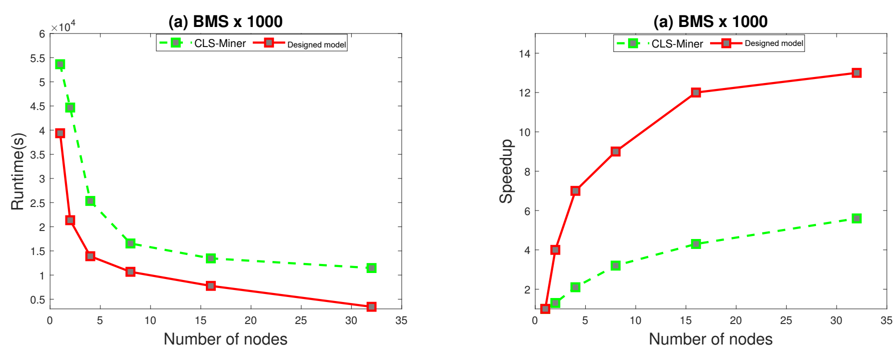


Figure 6: Scalability results.

the number of nodes varying from one to thirty-two. The results of this study show that the developed model has a significant advantage over the CLS-Miner in terms of both time required and speed of operation. Using a genetic algorithm and decomposition, it is possible to find closed patterns in large datasets that contain a significant amount of values. The created model is able to generate closed patterns that have high utility, and it has no problems managing huge datasets.

6. Conclusion

Due to the increasing number of devices connected to the Internet of Things, it is becoming increasingly difficult to manage huge datasets to identify high utility patterns that can be used for a variety of different industries and applications (IoT). Mining high utility transactional data used in market research is an important activity in market engineering as it can find valuable raw materials that enable efficient decision making. This is due to the fact that mining of high utility transaction data is used in market research. The process of developing marketplaces to meet the specific needs of individual customers is called market engineering. It is difficult to sift through all the patterns that can be discovered in a huge data set in a short period of time, but there are many that can be identified despite this challenge. Mining closed patterns that have high workloads is the main focus of current research, although the management of large datasets has received relatively little attention recently. Spark's three-tier MapReduce architecture was built in this paper that we could mine closed patterns in an effective manner despite the large amount of data that needed to be processed. In addition, a multi-objective k-means and a genetic algorithm (GA) are used to speed up finding patterns in the data. This can be done by narrowing the scope of the search so that only those options that have a higher probability of yielding desirable results are considered, rather than searching the entire search field. According to the results of a number of studies, the developed model is superior to the CLS-Miner in terms of speed, memory requirements, and scalability over a wide range of node counts. This was determined by comparing the performance of the two approaches.

7. Future Work

The results of the research conducted for this paper may be applicable to a variety of other areas of academic research. Although evolutionary computation was the main focus of this work, it would be beneficial to explore other techniques that could improve the proposed framework. Also, the datasets used in the study were already known in their respective domains. Nevertheless, it is possible that the HUIM approaches explored in this work could be used in other specific applications. Further research is needed if the tactics described in this study are to be implemented by these IoT networks as presented in this study.

References

[1] R. Agrawal, T. Imielinski, and A. N. Swami, Database mining:A performance perspective, *IEEE Transactions on Knowledge and Data Engineering*, vol. 5(6), pp. 914–925, 1993

[2] R. Agrawal and R. Srikant, Mining sequential patterns, *The International Conference on Data Engineering*, pp. 3–14, 1995.

[3] Y. Baek, U. Yun, H. Kim, J. Kim, B. Vo, T. Truong, and Z. H. Deng, Approximate high utility itemset mining in noisy environments, *Knowledge-Based Systems*, vol. 212, pp. 106596, 2021.

[4] Y. Chen and A. An, Approximate parallel high utility itemset mining, *Big Data Research*, vol. 6, pp. 26–42, 2016.

[5] J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters, *Communications of the ACM*, vol. 51(1), pp. 107–113, 2008.

[6] Y. Djenouri, J. C. W. Lin, K. Nørvåg, and H. Ramampiaro, ighly efficient pattern mining based on transaction decomposition, *IEEE International Conference on Data Engineering*, pp. 1646–1649, 2019.

[7] T. L. Dam, K. Li, P. Fournier-Viger, and Q. H. Duong, CLS-Miner: efficient and effective closed high-utility itemset mining, *Frontiers of Computer Science*, vol. 13, pp. 357–381, 2019.

[8] P. Fournier-Viger, C. W. Wu, S. Zida, and V. S. Tseng, FHM: Faster High-Utility Itemset Mining Using Estimated Utility Co-occurrence Pruning, *International Symposium on Methodologies for Intelligent Systems*, pp. 83–92, 2014.

[9] P. Fournier-Viger, J. C. W. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng, and H. T. Lam, The SPMF open-source data mining library version 2, *The European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 36–40, 2016.

[10] P. Fränti and S Sieranoja, How much can k-means be improved by using better initialization and repeats?, *Pattern Recognition*, vol. 93, pp. 95–112, 2019.

[11] P. Fournier-Viger, Z. Li, J. C. W. Lin, R. U. Kiran, and H. Fujita, Efficient algorithms to identify periodic patterns in multiple sequences, *Information Sciences*, vol. 489, pp. 205–226, 2019.

[12] R. Guha, M. Ghosh, S. Kapri, S. Shaw, S. Mutsuddi, V. Bhateja, and R. Sarkar, Deluge based genetic algorithm for feature selection, Evolutionary Intelligence, pp. 1–11, 2019.

[13] W. Gan, J. C. W. Lin, H. C. Chao, H. Fujita, and P. S. Yue, Correlated utility-based pattern mining, *Information Sciences*, vol. 504, pp. 470–486, 2019.

[14] J. H. Holland, Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, MIT Press, 1992.

[15] J. Han, J. Pei, Y. Yin, and R. Mao, Mining frequent patterns without candidate generation: A frequent-pattern tree approach, *Data Mining and Knowledge Discovery*, vol. 8(1), pp. 53–87, 2004.

[16] C. Lucchese and S. Orlando, and R. Perego, Fast and memory efficient mining of frequent closed itemsets, *IEEE Transactions on Knowledge and Data Engineering*, vol. 18(1), pp. 21–36, 2006.

[17] H. Li, Y. Wang, D. Zhang, M. Zhang and, E. Y. Chang, PFP: parallel fp-growth for query recommendation, *ACM conference on Recommender systems*, pp. 107–114, 2008.

[18] Y. Liu, W. Liao, and A. N. Choudhary, A two-phase algorithm for fast discovery of high utility itemsets, *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 689–695, 2005.

[19] J. C. W. Lin, T. Hong, and W. Lu, An effective tree structure for mining high utility itemsets, *Expert Systems with Applications*, vol. 38(6), pp. 7419-–7424, 2011.

[20] M. Liu and J. Qu, Mining high utility itemsets without candidate generation, *ACM International Conference on Information and Knowledge Management*, pp. 55–64, 2012.

[21] J. Liu, K. Wang, and B. C. M. Fung, Direct discovery of high utility itemsets without candidate generation, *IEEE International Conference on Data Mining*, pp. 984–989, 2012.

[22] M. Y. Lin, P. Y. Lee and, S. C. Hsueh, Apriori-based frequent itemset mining algorithms on MapReduce, *The International Conference on Ubiquitous Information Management and Communication*, pp. 1–8, 2012.

[23] Y. C. Lin, C. W. Wu, and V. S. Tseng, Mining high utility itemsets in big data, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 649–661, 2015

[24] J. C. W. Lin, W. Gan, P. Fournier-Viger, T. P. Hong, and V. S. Tseng, Efficient algorithms for mining high-utility itemsets in uncertain databases, *Knowledge-Based Systems*, vol. 96, pp. 171–187, 2016.

[25] J. C. W. Lin, L. Yang, P. Fournier-Viger, and T. P. Hong, Mining of skyline patterns by considering both frequent and utility constraints, *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 229–238, 2019.

[26] J. C. W. Lin, G. Srivastava, Y. Zhang, Y. Djenouri, and M. Aloqaily, Privacy preserving multi-objective sanitization model in 6G IoT environments, *IEEE Internet of Things Journal*, vol. 8(7), pp. 5340–5349, 2021.

[27] J. C. W. Lin, Y. Djenouri, G. Srivastava, U. Yun, and P. Fournier-Viger, A predictive GA-based model for closed high-utility itemset mining, *Applied Soft Computing*, vol. 108, pp. 107422, 2021.

[28] J. C. W. Lin, Y. Djenouri, and G. Srivastava, Efficient closed high-utility pattern fusion model in large-scale databases, *Information Fusion*, vol. 76, pp. 122–132, 2021.

[29] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Wu, DBSCAN revisited, revisited: why and how you should (still) use DBSCAN, *ACM Transactions on Database Systems*, vol. 42(3), pp. 1–21, 2017.

[30] G. Srivastava, J. C. W. Lin, M. Pirouz, Y. Li, and U. Yun, A pre-large weighted-fusion system of sensed high-utility patterns, *IEEE Sensors Journal*, 2020.

[31] Y. Djenouri and M. Comuzzi, Combining Apriori heuristic and bio-inspired algorithms for solving the frequent itemsets mining problem, Information Sciences, vol. 420, pp. 1–15, 2017.

[32] V. S. Tseng, B. Shie, C. Wu, and P. S. Yu, Efficient algorithms for mining high utility itemsets from transactional databases, *IEEE Transactions Knowledge and Data Engineering*, vol. 25(8), pp. 1772-–1786, 2013.

[33] V. S. Tseng and C. W. Wu and P. Fournier-Viger, and P. S. Yu, Efficient algorithms for mining the concise and lossless representation of high utility itemsets *IEEE Transactions on Knowledge and Data Engineering*, vol. 27(3), pp. 726–739, 2015.

[34] J. M. T. Wu, G. Srivastava, M. Wei, U. Yun, J. C. W. Lin, Fuzzy high-utility pattern mining in parallel and distributed Hadoop framework, Information Sciences, vol. 553, pp. 31–48, 2021.

[35] C. W. Wu, P. Fournier-Viger, J. Y. Gu, and V. S. Tseng, Mining closed+ high utility itemsets without candidate generation, *Conference on Technologies and Applications of Artificial Intelligence*, pp. 187–194, 2015.

[36] H. Yao, H. J. Hamilton, and C. J. Butz, A foundational approach to mining itemset utilities from databases, *SIAM International Conference on Data Mining*, pp. 482-–486, 2004.

[37] U. Yun, D. Kim, E. Yoon, and H. Fujita, Damped window based high average utility pattern mining over data streams, *Information Sciences*, vol. 144, pp. 188–205, 2018.

[38] U. Yun, H. Ryang, G. Lee, and H. Fujita, An efficient algorithm for mining high utility patterns from incremental databases with one database scan, *Knowledge-Based Systems*, vol. 124, pp. 188–206, 2017.

[39] M. J. Zaki and C. J. Hsiao, Efficient algorithms for mining closed itemsets and their lattice structure, *IEEE Transactions on Knowledge and Data Engineering*, vol. 17(4), pp. 462–478, 2005.

[40] S. Zida, P. Fournier-Viger, J. C. W. Lin, C. W. Wu, and V. S. Tseng, EFIM: a fast and memory efficient algorithm for high-utility itemset mining, *Knowledge and Information Systems*, vol. 51(2), pp. 595–625, 2017.

Mining of High Expected Utility Patterns from the Uncertain Databases

Usman Ahmed[1] and Jerry Chun-Wei Lin[1†]

[1]Department of Computer Science, Electrical Engineering, and Mathematical Science, Western Norway University of Applied Sciences, Inndalsveien 28, Bergen, 5063, Norway.

Contributing authors: Usman.Ahmed@hvl.no; jerrylin@ieee.org;
†Corresponding author

Abstract

Mobile or Internet of Things (IoT) devices have experienced phenomenal proliferation across a wide range of business sectors and fields of activity over the past few decades. As a direct result, an unimaginable amount of data is being produced and generated. The collected data contains a significant amount of interesting information (i.e., interestingness, weight, frequency, or uncertainty), but most of the existing and general algorithms for pattern mining consider only a single object and precise data to find the needed information. Due to the huge amount of information that is currently collected, it is very important to find information that is not only useful but also up-to-date within a certain time period. In this study, we take utility and uncertainty as the main objects to effectively identify the desirable High Expected Utility Patterns (HEUPs) in a limited time period based on a multi-objective evolutionary framework. Our goal is to complete this task with less time as possible. The ability of the developed model, known as MOEA-HEUPM, to find meaningful HEUPs in an uncertain environment without the use of specific criteria is one of the advantages of the model (i.e., minimal utility and minimal uncertainty). In developing HEUPM, two different schemas of coding were considered to for MOEA framework. Since the newly developed model MOEA-HEUPM allows the set of non-dominant HEUPs to be identified in a relatively short time, this speeds up the decision-making process. Experiments are then conducted to illustrate the utility and effectiveness of the newly developed MOEA-HEUPM model compared to generic approaches in terms of convergence, hypervolume, and the number of patterns discovered.

Keywords: data mining, uncertainty, utility, high expected utility pattern.

1 Introduction

Pattern mining algorithms have shown over the last decades that they are able to successfully identify important information for decision making [1]. The discovery of association rules that exist between different elements in databases can be done using a method called Apriori [2] , which is both the most basic and the simplest technique. It starts by defining the lowest level of support required, and then looks for the element sets that are most commonly used. The next step is to assemble a set of rules by combining the satisfied frequent sets of objects. After that, the confidence of an association rule is considered as a rule if it is greater than a minimum confidence level. It has been proven that association rule mining (ARM) is an effective knowledge discovery technique as it is used in a variety of different domains and applications. ARM, on the other hand, only considers the binary database and does not consider additional features such as weight, importance, interestingness, or quantity, among others. Basket-market analysis is an illustrated example of this concept. ARM does not take into account items included in a transaction and their purchase amounts, even if they are included in the transaction. Another disadvantage of ARM is that each item is evaluated and ranked according to the same importance scale. In a transactional database, ARM has been used to determine the relationships between each itemset. However, because ARM has its limitations, it is possible that important and beneficial patterns may be overlooked, leading to inappropriate decision-making techniques. Although ARM has been used to detect the relationship between itemsets in a transactional database, it has limitations that prevent it from detecting all patterns.

To obtain the set of high-utility itemsets (HUIs) from quantitative datasets, a new field of study known as high-utility itemset mining (HUIM) [3, 4] considers both item unit profit and its quantity. The main goal of the High-Utility Itemset Mining (HUIM) is to identify patterns that are of high utility to users. If the utility of an itemset is greater than a pre-determined minimum utility threshold, that itemset is called a High-Utility Itemset (HUI). Thus, based on HUIM, it is possible to find more profitable itemsets that can be used to make more successful strategic decisions. These can be used in a variety of contexts. In addition, most of the currently available techniques for pattern mining also depend on a priori criteria to figure out what information is needed for investigation [1]. This is not a simple task, as it requires technical and domain knowledge to find an appropriate threshold that avoids difficulties such as “rare elements” and “combinatorial explosion”. This is not an easy task, because it requires technical and professional domain knowledge, which also arises many studies in recent years.

Processing data in an environment as complicated as industry presents a number of challenges. One of them is the lack of understanding of the data sources and the many environmental components. Because many resources (i.e., WiFi systems, RFID tags, wireless sensor networks, and GPS) are subject to uncertainty, traditional data mining techniques are unable to extract all the necessary information from uncertain databases. This is because conventional data mining techniques rely on resources that are not subject to uncertainty. Utility and uncertainty are two different measures that provide semantic and objective value for each pattern. For this reason, we believe that these two components are essential. The utility of the pattern, which is a semantic mechanism, is used in traditional data mining techniques as a measure to assess how useful the pattern is. These approaches depend on the usefulness of the pattern. In other words, the usefulness of a pattern depends not only on whether or not it is present in the data at hand, but also on whether or not it is useful. One can think of the uncertainty measure as an objective measure that assesses the reliability and existence of the pattern in terms of the probability that it is actually present. These two facets are in no way equivalent. When mining HUIs from uncertain datasets, it can be difficult to consider the utility components and the uncertainty features simultaneously. In the vast majority of cases where utility-based approaches are used, it is assumed that the data sources are correct; however, the uncertainty component of the data is not considered. If the uncertainty factor is not taken into account, it is likely that the extracted patterns will become unusable and unreliable, and that they will lack essential information. It is also possible that they are missing crucial information. The probability of this being the case is very low. Moreover, more than two aspects are considered at the same time when making a decision, and these considerations might compete with each other or not be the most important ones (e.g., price and distance to the city center when booking a hotel room). To obtain the information needed to make an effective decision in a short time, a trustworthy method that takes into account both the uncertainty factors and the utility variables is essential.

In this study, we first take uncertainty and utility as the two objects of consideration to find the non-dominant solutions based on evolutionary computation. Then we discuss the contributions made by this paper as follows.

- In this paper, we investigate the issue by simultaneously considering utility objects and uncertainty objects or multi-objects to discover qualified non-dominated patterns with high expected utility (HEUPs) from uncertain datasets using evolutionary computing.
- It is decided to design a model known as a multi-objective evolutionary approach for mining patterns with high expected utility (MOEA-HEUPM). This approach can be used to determine the necessary HEUPs for an uncertain environment in a very short time.
- The more advanced MOEA-HEUPM does not require prior knowledge, also known as minimum utility threshold or minimum uncertainty threshold, to find new information. Instead, it is possible to identify the non-dominated

patterns, a method that is much more unique and helpful in the process of problem solving and decision making.

- In this situation, the weight-based Tchebycheff approach is used so that non-dominated solutions can be generated immediately based on the multi-criteria weighting mechanism regarding utility and uncertainty.
- Experiments have shown that the resulting MOEA-HEUPM performs better than generic pattern mining algorithms in terms of convergence, hypervolume, and the amount of patterns identified. Subsequently, two different coding schemes were created and used in the model MOEA-HEUPM, which was created to illustrate the usefulness of the developed model.

2 Literature Review

Mining association rules (ARM) is the most important information in the field of Knowledge Discovery in Databases (KDD), and the first method for identifying connections between items in databases was called Apriori [2]. Agrawal and Srikant [2] are responsible for the development of this method. Apriori uses a two-stage procedure to determine which element groupings are most common in the database in the first stage. Then, the combinations of the most frequently occurring components are formed so that they can serve as the basis for the association rules, which are based on the minimum confidence level. To manage the ARM in terms of frequency of occurrence for the efficiency performance, a large number of extensions [1, 5] have been fully provided and studied. In a real-world application, the number of diamonds purchased in a shopping mall is much less than the number of clothes; however, the profit the retailer makes on the diamonds is much greater than that on the clothes. This is because the frequency of occurrence is not indicative of the patterns discovered. The reason is that the frequency of occurrence does not indicate anything about the insight into the patterns found. The term "high-utility itemset mining" or HUIM refers to a method developed in the last few decades. It was first introduced as a method for extracting useful itemsets from databases [4, 6]. To determine whether a database contains high-utility itemsets (HUIs) or not, HUIM considers not only the single gain of items (called profit) but also the total number of items (quantity) of the transactions in the database. Since the original HUIM does not include the downward closure property, the two-stage model known as transaction-weighted utility (TWU) was introduced to construct the downward closure property by using high-transaction-weighted utility itemsets [6]. For this reason, the generated HUIs should continue to be accurate and complete. Most algorithms for pattern mining [7], such as ARM and HUIM, require the development of a minimum threshold to judge whether or not an item or collection is considered a useful, relevant pattern. Several other extensions were considered and one of the strategies examined for its ability to find HUIs without the need for thresholds was called top-k HUIM [8]. In determining the set of HUIs, an evolutionary calculation was used as a common measure in addition to the traditional minimum threshold. In this way, information from databases could be uncovered, which is not only important but also helpful.

Kannimuthu et al. created a GA-based model with a ranking mutation operator to detect HUIs in their research [9]. However, the model built on GA requires some work in processing. To search the HUI collection in a different way, Lin et al. proposed the use of a PSO-based model [10]. In addition, an effective OR/NOR-tree [11] was developed to check the real solutions in the evolutionary development process, which can lead to more accurate HUIs. This was done to guarantee that the evolutionary process would produce accurate results, and it was successful. After that, the ACO-based method was also developed, which was named HUIM-ACS [12] and was developed to find the HUI set efficiently. On the other hand, previous works focused on either finding frequent object sets or finding object sets with high utility using a priori parametric factors (i.e., minimum support threshold or minimum utility threshold). It is not possible to use the techniques described above to find the intriguing patterns that involve more than one component, such as uncertainty and utility interacting with each other. Because the Internet of Things (IoT) is growing so rapidly, many types of data are being collected from a variety of unpredictable environments. Therefore, it is essential to develop a reliable strategy to deal with this problem especially by collecting the uncertain data in the IoT environments.

Generic evolutionary computation (EC) [13] is a meta-heuristic technique used to address NP-hard and optimization problems in an efficient way. This approach is based on the development of a single-objective fitness function during the evolutionary process. This strategy was developed to address problems that could not be solved using conventional approaches. The Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Ant-Colony Optimization (ACO) are the most widely used methods EC [14]. These methods are based on the idea that there is only one fitness function that is important in the course of evolutionary development. According to [13, 15], multi-objective optimization is one of the most studied areas of evolutionary computing and is used in a variety of industries and applications. The MOEA/D framework is an example of a broad framework that divides a larger multi-objective evolutionary problem into a set of more manageable multi-objective optimization subproblems. To find optimal solutions to each of these individual problems, MOEA/D uses a technique called population-based optimization [16]. It is able to provide a collection of solutions that are uniformly distributed and shows significant convergence compared to other MOEA algorithms such as NSGA-II [17] and SPEA-II [18].

In the field of data mining, multi-objective evolutionary algorithms, also called MOEAs, are often used. This is because classification, clustering, and feature selection are all considered to be in need of improvement. This is due to the fact that questions that need to be optimized fall into the category of multi-objective [13]. A variety of criteria, such as interestingness, support, trustworthiness, understandability, and lift, are used to identify, for example, the most intriguing pattern in the database. Zhang et al. have developed an evolutionary approach used to search for common and valuable patterns [19]. This model does not require specific parameters, but discovers the patterns that are not dominated by the components that are helpful and frequent. On the other hand, this approach does not uncover enough patterns on which to base judgments, so it is not suitable for this purpose. Djenouri et al. [20] have developed a set of meta-heuristic algorithms to efficiently search the datasets for patterns that have high frequency and usefulness. Mining HUIs from an uncertain database is the main focus of models that have not yet been developed; therefore, it is the main concern of this work.

In recent years, a variety of algorithms for mining uncertain patterns have been developed to extract valuable patterns from databases. This goal has been the focus of research and development in this area. High-utility uncertain itemsets [21], frequent uncertain patterns (UFPs) [22, 23], uncertain sequential patterns [24, 25], uncertain weighted frequent itemsets [26], and interesting uncertain patterns [27] are examples of these types of patterns. The method proposed by Liu et al. [24] for extracting uncertain sequential patterns is based on the candidate generation approach. This method was then applied to the sensor dataset obtained from the pollution monitoring network. Palacios et al. [25] used the fuzzy uncertainty mining technique to identify the events that were present in the uncertain health data from the aviation engine status monitoring. Potential High-Utility Itemsets Mining is proposed by Lin et al. [21] to the novel approach they developed for the problem of uncertain databases (PHUIM). The uncertainty tuple model is implemented in PHUIM to find uncertain patterns with high utility. Lin et al. [26] then proposed an Apriori-like two-phase technique to identify weighted frequent items. They also advocated a high upper-bound to reduce the size of the search space and the number of irrelevant item sets. This work serves as the baseline for the uncertain closed itemsets with high utility cited by Bui et al. [28]. The model uses the downward closure property and a depth-first search to obtain the needed information. Both search methods are used to eliminate unclosed potentially valuable itemsets.

In their study, Ahmed et al. developed a weighted technique for identifying potentially interesting patterns in ambiguous data [27]. In this method, the values of prefixes are found using a structure built on trees. Then, the weight-based technique developed by Lee et al. [23] for mining ambiguous patterns was used to select the relevant groupings of objects. A list-based data format that was used to mine the frequent uncertain patterns [22]. Yun et al. [29] pointed out that the performance problem of the current incremental high-utility mining approach, which has a high false discovery rate and generates a large number of patterns. Then, the list-based mining technique was described to discover high-utility patterns during incremental mining without using candidate generations. In addition, Yun et al. proposed a window-based technique for extracting utility patterns from data streams [30]. The algorithm eliminates the need for a large number of candidates that have a low probability of success to go through the creation and testing process, and it is able to perform well in complex dynamic systems. In short, most of the study focused on isolating patterns from an unreliable database that stood out for either their frequency or their utility, using two different metrics known as support and utility factors. These two metrics were used by a smaller number of search techniques to find the critical information that occurred less frequently. The TKQ-Miner [31] algorithm considered two factors and used constrained estimation when searching for the top-k patterns with the highest utility. On the other hand, the use of TKQ-Miner required the use of priori parameters, which included a minimum utility value in addition to a minimum support value. In the real-world applications, users must have a certain level of knowledge to create the appropriate configurations for mining patterns with high utility value.

Mai et al. [32] investigated the issue of increasing mining productivity by focusing on a small number of non-redundant itemsets with high utility. Using a lattice structure, the newly developed approach NR-HARs mines a smaller collection of HUIM than was previously possible. This strategy facilitates timely decisions by examining a limited but carefully selected set of high utility items. Baek et al. [33] came up with the idea for the method known as uncertain itemset mining, a strategy based on the use of lists. This method selects the items with the highest values, which in turn helps to find items produced at manufacturing sites that are free of defects [33]. Lee et al. [34] created a tree structure for the data using a technique known as uncertain frequent pattern mining. The approach allows to explore uncertain data and helps to overcome the limitations of traditional techniques that are not able to deal with the probabilities of individual components during the pattern mining process. The problem of an uncertain HUIM was addressed by Lin et al. [35] who considered the availability of probabilistic models of transactions based on a list structure. By using efficient pruning algorithms, the developed method can avoid a large number of database scans and eliminate a significant number of records of items that prove to be unpromising at an early stage. Subsequently, Gan et al. [36] developed the HUPNU algorithm (Mining High-Utility itemsets with both Positive and Negative unit profits from uncertain databases), which considers both the positive and negative HUIs present in uncertain databases. To calculate the HUIM using the techniques described above (which may vary depending on the dataset), you must have established criteria. On the other hand, the proposed technique does not require thresholds to figure out how to extract the high utility itemset from the uncertain database. On the other hand, during the pattern mining process, it considers both the unpredictability of the results and the utility of the data as multi-objectives.

In this work, the created MOEA-HEUPM model is used to extract patterns with high expected utility from uncertain datasets using the MOEA [16] algorithm. The MOEA algorithm used in building the model has the advantage of providing a set of non-dominated solutions, which is one of its advantages (it represents the best trade-off relationships between multi-objects, i.e., utility and uncertainty). Thus, users can choose the solutions that are best suited for their decision making, according to their own tastes. Moreover, the number of solutions that can be found is limited, and non-dominating solutions can be found within this time limit; therefore, the choice or strategy for real-time applications can be made effectively.

3 Preliminaries and Problem statements

Let there be a collection of items denoted by $I = \{\imath_1, i_2, \ldots, i_m\}$, and let there be an uncertain transaction database represented by a set of transactions denoted by $D = \{T_1, T_2, \ldots, T_n\}$, where each transaction is a set of items that can be converted to expressions such as $uv(i_k, T_c)$ can be used to specify the value of each component of a transaction that has a value and can be affected by uncertainty. The Table 1 provides a simple example of the uncertain and quantitative database. Table 2 also shows the individual unit profits for each product included in the database.

Table 1: The quantitative and uncertain database

TID	**Item: quantity, probability**
T_1	(a:5, 0.3); (b:3, 0.40); (c:6, 0.9)
T_2	(c:4, 0.75); (d:2, 0.9)
T_3	(a:7, 1.0); (b:8, 1.0); (e:2, 0.75)
T_4	(a:3, 0.9); (c:1, 0.9)
T_5	(b:2, 1.0); (c:4, 0.95); (e:4, 1.0)

Table 2: The unit profit of the items

Item	**Profit**
a	8
b	3
c	8
d	3
e	5

A recurring example in this work is the Tables 1 and 2. It is then described below: it consists of five transactions, i.e., $(T_1, T_2, T_3, T_4, T_5)$, respectively. For example, in transaction T_2, the items (c) and (d) were sold. The purchase quantities for these items are 4 and 2, respectively, and their unknown values in transaction T_2 are 0.75 and 0.9. The unit profit for each item sold can be found in the database in the Table 2. For example, if one item was sold, the retailer would receive 8 as profit per unit of an item (a).

Definition 1 (Item Utility in a Transaction) The term $u(i_k, T_c)$, defined as follows, denotes the utility of an item with a value of i_k when considered in the context of a transaction with a value of T_c.

$$u(i_k, T_c) = pr(i_k) \times q(i_k, T_c) \tag{1}$$

Example 1 For example, the calculation for evaluating the usefulness of an item (a) in the context of the variable T_1 is as follows: $u(a, T_1) = 5 \times \$8 = \$40$.

Definition 2 (Itemset Utility in a Transaction) The $u(X, T_c)$ represents the utility of an itemset, denoted by the X, in a transaction denoted by T_c. This expression is satisfied if $i_k \in X \subseteq T_c$. It can be characterized as follows:

$$u(X, T_c) = \sum_{ik \in X} u(i_k, T_c) \tag{2}$$

Example 2 For instance, the utility of an itemset (ab) in T_1 can thus be computed as follows: $u(ab, T_1) = \$40 + \$9 = \$49$.

Definition 3 (Itemset Utility in a Database) In a transaction database D, the utility of X is symbolized by the symbol $u(X)$, which has the following definition:

$$u(X) = \sum_{X \subseteq T_c \wedge T_c \epsilon D} u(X, T_c) \tag{3}$$

Example 3 For example, the utility of an itemset (ac) in T_1 is calculated as: $u(ac) = u(a, T_1) + u(c, T_1) + u(a, T_4) + u(c, T_4) = \$40 + \$48 + \$24 + \$8 = \$120$.

Definition 4 (Itemset Uncertainty in a Transaction) The uncertain value of an itemset, denoted X, is represented in a transaction by the symbol $uc(X, T_c)$, and its definition is as follows:

$$p(X, T_c) = \Pi_{x_i \epsilon X} p(x_i, T_c) \tag{4}$$

Example 4 For example, the probability that the value (a) is present in the set T_1 can be calculated as follows: $p(a, T_1) = 0.3$. The probability that the value (ac) is present in the set T_1 can be calculated as follows: $p(ac, T_1) = p(a, T_1) \times p(c, T_1)$, which equals 0.3 multiplied by 0.9, which equals 0.27.

When it comes to generic pattern recognition approaches, most of the algorithms in use today focus on analyzing patterns with respect to specific thresholds such as support, confidence, uncertainty, or utility [1, 5]. Moreover, choosing the appropriate threshold for evaluating patterns is not an easy task, as it can quickly lead to problems such as "rare-item" and "combinational explosion" [19]. This makes the task a difficult one. For example, the number of patterns that can be detected decreases when the threshold is raised to a higher level. On the other hand, a significant amount of additional data is revealed when the threshold is raised to a higher level. Pattern mining requires a priori knowledge to determine what is an acceptable threshold. It is not a simple process to find the patterns that are most useful for decision making. In addition, most KDD algorithms are designed to collect information from various databases. The information obtained then still has the value of uncertainty in the databases, in this era of the Internet of Things (IoT). The limitations mentioned earlier will be addressed in the following discussion when we state the problem in this paper.

Problem Statement: This study focuses on the analysis of both utility and uncertainty elements in the quantitative and uncertain database to identify the set of non-dominated patterns with high predicted utility without the need to pre-specify the knowledge discovery criteria. More specifically, this goal is achieved by using a pattern discovery technique that does not depend on a priori thresholds. In this situation, evolutionary multi-objective computation, also known as MOEA, is used to obtain the necessary information in a limited period of time. This helps to reduce the computational cost associated with processing the large dataset and determining the most up-to-date information.

4 Developed MOEA-HEUPM Model

In the next part, we first introduce the MOEA-based model, which we named MOEA-HEUPM, to determine how to extract the non-dominated High Expected Utility Patterns (HEUPs) from quantitative and uncertain datasets. For an explanation of the proposed model, see the Fig. 1 of the designed approach.

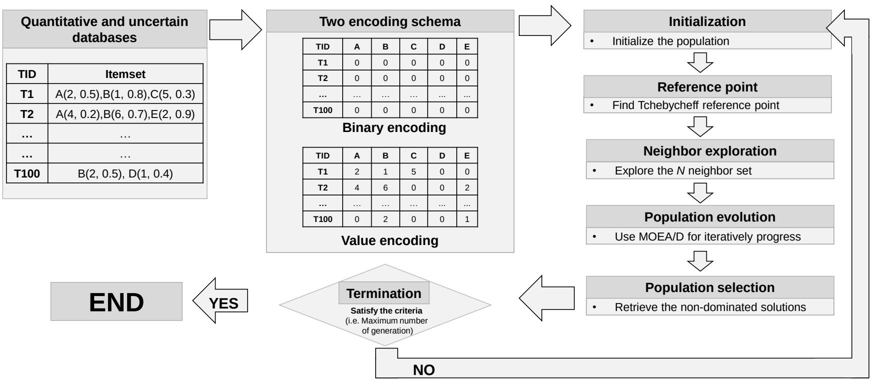


Fig. 1: The framework of the designed MOEA-HEUPM model

The MOEA-HEUPM was developed with the intention of finding the collection of non-dominated High Expected Utility Patterns (HEUPs) in the uncertain datasets. This was done with the aim of considering both the utility and the uncertain aspects of the situation. The two different measures cannot be used interchangeably due to a number of incompatibilities. In certain situations, the patterns that have a higher value do not also have a higher degree of uncertainty, and the patterns that have a lower utility do not lead to a higher utility when combined. The scenario can be understood as a two-objective optimization problem to analyze these two aspects simultaneously without relying on prior knowledge. Compared to more traditional pattern mining methods, the developed model reflects only a small number of patterns that are not only practical but also relevant to the process of decision making. Due to the fact that utility and uncertainty are considered in the developed model MOEA-HEUPM, two objectives are considered maximized, which are represented by the Eq. (5). The details of the proposed structure are explained in the following sections.

$$Max\,F(X) = \left\{max\,(utility(X), uncertainty\,(X)^T\right\} \tag{5}$$

4.1 Initialization

The phase of the multi-objective evolutionary strategy known as **initialization of the population** is an important phase. This is because the population may lead to a solution that is not optimal, or it may require a significant amount of computational resources to converge [13]. In most cases, the initial population was generated by randomly selecting elements from the databases, which eventually led to the randomization of the population. Nevertheless, the random selection method will always result in spurious patterns. That is, the pattern is not contained in the transactions themselves, where it should be looked for. The first step in the development of MOEA-HEUPM was to present the problem-specific initialization approach. In selecting the members of the population, we use two different procedures, one called **transaction-itemset-selection** and the other called **meta-itemset-selection**. In the technique known as **transaction-itemset-selection**, the fifty percent of individuals that make up a population are selected from the transactions stored in the database and prioritized according to the likelihood that they will be of use. This technique ensures that each selected individual comes from the pool of individuals with the highest quality responses in the database. The uncertainty probability for the remaining fifty percent of the individuals that make up a population of **meta-itemset-selection** is used to code and select only one item from a transaction at a time.

When it comes to the evolutionary process, the transaction itemset can be helpful in the timely search for optimal solutions (i.e., crossover and mutation operations). Since the transaction itemset is selected based on the utility probability, it is possible that the solutions converge quickly and reliably. On the other hand, the meta-itemset can be useful when trying to generate new offspring from parents that have a greater variety of traits. Since the meta-itemset consists of single elements, it is more likely to generate new offspring when exposed to the crossover operator. Therefore, it is possible to get a wider range of solutions when you use it. Therefore, the very first thing we do in this phase is calculate the probabilities associated with the meta-itemset (Algorithm 1, lines 1 to 3). The results are then displayed in Table 3. Next, we calculate the utility probability of the itemset being traded (Algorithm 1, rows 4 to 6). The results are displayed in Table 4. The population is formed by first randomly selecting individuals from Tables 3 and 4 and then randomly generating the remaining fifty percent of these individuals. This process is called "random selection" (Algorithm 1, line 7). Therefore, in the first generation, the population will initially consist of only four individuals. The created MOEA-HEUPM will randomly select two individuals from Table 4 based on the utility probability and randomly select two individuals from Table 3 based on the uncertain probability. This selection was made in an appropriate way. The algorithm used for initialization can be found in the file named Algorithm 1.

Table 3: Meta-itemset-selection strategy

Itemset	Uncertainty	Uncertain Probability
a	1.3	0.16
b	1.4	0.17
c	2.75	0.34
d	0.9	0.11
e	1.75	0.22

Table 4: Transaction-itemset-selection strategy

TID	a	b	c	d	e	Utility	Utility Probability
$T1$	5	3	6	0	0	97	0.31
$T2$	0	0	4	2	0	38	0.12
$T3$	7	8	0	0	2	90	0.29
$T4$	3	0	1	0	0	32	0.10
$T5$	0	2	4	0	4	58	0.18

Algorithm 1 MOEA-HEUPM: initialization

INPUT: The transaction data set D, population pop size, weight vector $\{w_1, w_2, \ldots, w_{pop}\}$, the size of neighbors n_s, crossover probability p_c, mutation probability p_m.

OUTPUT: Initialize individuals in a population.

1: **foreach** $i \in meta_{itemset}$ **do**

2: $M \leftarrow (\frac{uncertain(D_i)}{uncertain(D)})$;

3: **end foreach**

4: **foreach** $j \in transaction_{itemset}$ **do**

5: $T \leftarrow (\frac{utility(D_j)}{utility(D)})$;

6: **end foreach**$P \leftarrow initial(M, T, pop)$;

7: **Return** P.

4.1.1 Encoding

In the field of evolutionary computing, researchers have experimented with a variety of coding strategies, each corresponding to their own unique application areas and domains. However, in EC, the binary coding technique and the value coding method are commonly used because they can contribute to higher convergence and greater diversity in the solutions generated [13]. After initializing the population in the model we constructed, the next step was to code the initialized population (Algorithm 2, line 1). To begin with, the model we developed MOEA-HEUPM uses not one but two different encoding methods, namely binary encoding and value encoding. If an item is part of a transaction, the binary coding scheme assigns a value of 1 to the associated item, and if it is not part of the transaction, the item is assigned a value of 0. Once it is determined that the value will be used for value coding, the set of items is used as the coding value in the scheme. According to [37], the fitness function for a given encoding scheme depends on two factors: the value and the order of the values. In the next section, we elaborate on these properties. The form of a scheme that remembers only the order is called a permutation and is the type of scheme most commonly used to order questions cite66. Binary and value encoding schemes, on the other hand, remember both the order and the value of the data. It is possible to calculate the value of an item by multiplying the quantity of the item by its price. This applies to each individual item that is part of the transaction. The value of uncertainty can also be used as a coding value. The uncertain value, on the other hand, is not directly related because it varies from item to item and contributes to the total uncertain value of the transaction. This contributed to the transaction having an uncertain value.

4.1.2 Reference point

Within the developed MOEA-HEUPM, the Tchebycheff reference point is used to find the solutions that dominate the noon hours based on the utility and uncertainty components (Algorithm 2, line 2). The Tchebycheff value is used to assess how trustworthy the coding scheme [19] given in Eq. (6). The purpose of the Tchebycheff value is to provide an efficient set that makes it easier to find non-dominated patterns that have high predicted utility.

$$\texttt{min}\ g^{te}(X\|w_i, z^*) \leftarrow \texttt{max}_{j=1}^{2}\{w_i^j \cdot (\|F_j(X) - z^*\|)\} \tag{6}$$

The procedures for determining the reference point are broken down and illustrated with an example further down the page. Assume that the number of subproblems is already stored in the variable *pop*, and define W as a set of even weights with the following format: $(w_1, w_2, \ldots, w_n)$, where $w_i^1 + w_i^2 = 1$. Imagine we have an individual with the following coding: 1, 0, 1, 0, 0 with a total utility value of 22 and a corresponding uncertainty value of 0.13. Suppose that the maximum possible imprecision of the database is evaluated as 0.715, while the largest potential utility of the database is evaluated as 107. The largest combination of utility and uncertainty value is assumed to be the individual reference point $z*$. This is done in accordance with the Eq. (6) (Algorithm 2, lines 3 to 5). Assume that the uncertainty weight value is 0.34 and the utility weight value is 0.66. These numbers indicate the even weights assigned to the uncertainty and utility categories, respectively. Next, we calculate the utility value of the individual based on the Eq. (6), which is $0.34 \times$22-107 $= 28.9$, and we calculate the uncertainty value of the individual as $0.66 \times$ 0.13-0.715 $= 0.38$. Both values are expressed in dollars. Next, the value of Tchebycheff is determined by calculating the largest difference between 28.9 and 0.38, which is 28.9. When this step is completed, the Tchebycheff values for each person in the population are calculated and evaluated.

4.1.3 Neighbor exploration

Next, we decide who the immediate neighbors of the people to the left and right of them are (Algorithm 2, lines 3 to 5). The *pop* is the value computed for each weight vector w_i ($i < pop$), where *pop* is the population size initialized at the beginning of the process. The formula for calculating the Euclidean distance between two individuals in a population (*pop*) is as follows: $\sqrt{(\|u(ind) - w_i^1\|^2 + \|p(ind) - w_i^2\|^2)}$. The term "individual Euclidean distance" is used to refer to this particular distance. As the population evolves, a mutant child is generated that requires a two-way crossover operation in the neighboring group. The next algorithm, referred to as Algorithm 2, provides a more detailed breakdown of the calculation for the reference point as well as the examination of the neighbors.

Algorithm 2 MOEA-HEUPM: reference point calculation and neighbor exploration

INPUT: P, the size of neighbors n_s, crossover probability p_c and mutation probability p_m.
OUTPUT: Population with neighbors. $P \leftarrow Encode(P)$ ▷ Binary or Value encoding $z^* \leftarrow$ initialize reference point.
1: **for** *all* $p \in pop$ **do**
2: $N_i \leftarrow$from P get the ns individual using Euclidean distance between any individual in P the weight vector W_i;
3: **end for**
4: **Return** N_i.

4.2 Population evolution

In the rest of the development process, the MOEA/D technique of the constructed MOEA-HEUPM model is implemented here (Algorithm 3, lines 1 to 6). For each individual who is already a member of Pop_i, an individual from N_j, assumed to be a neighbor of Pop_i, is randomly selected to join $Pop_i^{'}$. Then an operator that performs a crossover in both directions is applied to the relationship between the variables Pop_i and $Pop_i^{'}$. The mutation operator is also used in the same way as everything else. After that, the Tchebyshev value of the child is determined and then compared to the Tchebyshev value of the neighbor N. If the individual's Tchebyshev value is of better quality, it is used to replace the Pop_i with the offspring and update the reference point z^* as previously indicated (Algorithm 3, lines 2 to 5). Then this process is repeated as many times as necessary until the termination condition max_{gen} is satisfied. The *Tchebyshev value* [19, 38] provides an alternative technique for finding the solutions that are not dominated by other variables. This allows us to avoid the standard approach used in pattern mining with the criteria already established (i.e., minimum utility and minimum uncertainty).

Algorithm 3 MOEA-HEUPM: population evolution

INPUT: P, N_i and number of generations max_{gen}
OUTPUT: Non-dominated solutions

1: **while** max_{gen} **do**
2:     **for** *all* $i \in pop$ **do**
3:         $P_i^{'} \leftarrow Randomly\ select\ an\ individual from\ N_i$;
4:         $child \leftarrow CrossMutation(P_i^{'}, P_i)$ ▷ Compute *child* objective function, if the Chebyshev value of the child is better than any individual *ind* in N_i, then replace *ind* with child and update reference point z^*
5:     **end for**
6: **end while**
7: $Final_{solution} \leftarrow SelectNonDominatedItemsets(P)$; ▷ Apply fast non-dominated sorting strategy to get the non-dominated item sets from final population P
8: **Return** $Final_{solution}$.

4.3 Population selection

When the MOEA/D process, which is constantly changing, is finally completed, the final population is generated and produced. Then, the individuals that make up the population are sorted and selected to represent the non-dominant responses for the final results [39] (Algorithm 3, lines 7 to 8). To illustrate the phases, we take the case of population Pop with K fronts. This population has a total of $(1 < i < K)$. In the first step of the process, the Pop function is used to find all non-dominant solutions, each of which is then assigned to the F_1 variable. The values previously set to $F1$ are then removed, leaving behind the value $Pop - F_1$. In this way, all possible fronts can be recovered and considered for inclusion in the final plan. The information in Fig. 1 shows that the evaluation is complete when the algorithm reaches the maximum number of generations allowed. After this step, the solutions are developed and made available to users. The subsequent steps of the algorithm are described in detail in the previously Algorithm 3.

5 Experimental Evaluation

In this section, we compared the proposed model MOEA-HEUPM with two baseline algorithms. The U-Apriori algorithm [40] and the EFIM algorithm [41] are examples of basic computational methods. Both the U-Apriori method and the EFIM algorithm require the minimum threshold for uncertainty, but the EFIM algorithm also requires the minimum threshold for utility. Evaluations were performed on a PC running Windows 10 equipped with an AMD Ryzen 5 PRO 3500U central processing unit and 16 gigabytes of memory (RAM). For MOEA-HEUPM, the population size was set to 100, the maximum generation was also set to 100, the neighborhood size was set to 10, the crossover probability was set to 1.0, and the mutation probability was set to 0.001.

Table 5: The conducted datasets for the experiments

Dataset	#Transactions	#items	Ave-Length	Type
Chess	3196	76	37	Dense
Mushroom	8124	120	23	Dense
Pumsb	49,045	21113	74	Dense
Kosarak	9899	1044	8	Sparse
Retail	8162	8345	10	Sparse
Accident	34018	468	33	Dense

In order to carry out the experiments, we acquire six datasets that vary greatly in quality from one another. Each one of them may be accessed individually using the SPMF library [42]. Table 5 outlines the properties of the datasets for your perusal. These include the number of highlighted transactions (highlighted with the symbol **#transaction**), the number of highlighted line items (highlighted with the symbol **#item**), the average length of transactions in the highlighted dataset (highlighted with the symbol **Ave-length**), and the type of dataset (highlighted with the symbol **type**) (i.e., dense or sparse). The chess dataset includes 3,196 transactions, sometimes known as moves, and 76 unique objects. Moves are another term for transactions (position on the chess board). The mushroom dataset has 8,124 transactions, each of which is an average of 23,120 items. The information included in the pumsb is summed up in the population and residential density. Pumsb dataset typically has a length of 74 and include 49,046 transactions and 7,116 elements that are exclusive to themselves. The kosarak dataset is data from a Hungarian news site named kosarak. It includes details regarding transactions completed through clickstream. It is comprised of 1,044 different types of commodities and 9,899 separate transactions, with each one lasting an average of eight minutes. The retail dataset was supplemented with 8,162 client interactions by a Belgian merchant who requests to remain nameless. There are 8,345 distinct items now available for purchase, and the typical transaction takes 10 seconds to complete. The accident dataset is a collection of incidents of vehicular collision that has place in Belgium (the anonymous incident occurred on a public road). There were 34,018 transactions, and the average number of items that were purchased in a single transaction was 33. To then determine the utility and uncertainty values for the datasets, we used a technique quite similar to the one we had previously explored [7]. The individual item probabilities were then randomly assigned to the itemset of each transaction within the range of 0.0 to 1.0 according to the HUPM [21] normal distribution. To evaluate the patterns found, we use two established scales, referred to separately as hypervolume (HV) and coverage (Cov) [19]. The HV is used to evaluate the dispersion and convergence of element groups that are within the target's search region, which in our particular scenario would be the values for uncertainty and utility.

The "hypervolume" (HV) can be described as follows, according to [19, 38]: Let A be defined as $(x_1, x_2, \ldots, x_l) \subseteq X$, where X is a set consisting of l decision vectors. What the function $S(A)$ returns is the volume enclosed by the union of the polytopes $p_1, p_2, \ldots, pl$, and each p_i is generated by the intersections of the following hyperplanes resulting from x_i and the axes: There is a hyperplane perpendicular to each axis in the target space and passing through the location where the axis is. If you are working with only two dimensions, each pi represents a rectangle formed by the points $(0, 0)$ and $(f_1(x_i), f_2(x_i))$ [19]. The Cov value is a common measure used to illustrate the different types of solutions found and the different behaviors of the itemset [19]. When the Cov value is large enough, the derived solutions have a greater variety of items, and the non-dominated solution set has its own set of items that is different from the other sets. The convergence is represented by Eq. (7).

$$Cov = N_d/N, \tag{7}$$

where N is the total number of discovered patterns and N_d is the different patterns.

5.1 Encoding schema analysis

Following this step, the encoding strategies, also referred to as value encoding and binary encoding, are compared in terms of multiple generations for both the Cov and HV metrics. Then, HV and Cov are put through their paces to determine whether or not the two different binary and value coding schemes are successful. The results for the two independent versions of the coding schemes are then shown sequentially in Figs. 2 and 3. According to the data shown in Fig. 2, value coding converges after a very small number of generations, while binary coding yields better results after a larger number of generations, especially when applied to a sparse dataset. This is because binary coding uses more bits per generation. Therefore, we can conclude that binary coding is superior for datasets with a considerable number of itemsets. This is the conclusion we can draw. On the other hand, the use of a value coding scheme should be considered when the total number of itemsets in the dataset is relatively small or when the dataset is of the dense type. The behavior of HV is shown in Fig. 3 and is similar to the behavior of Cov.

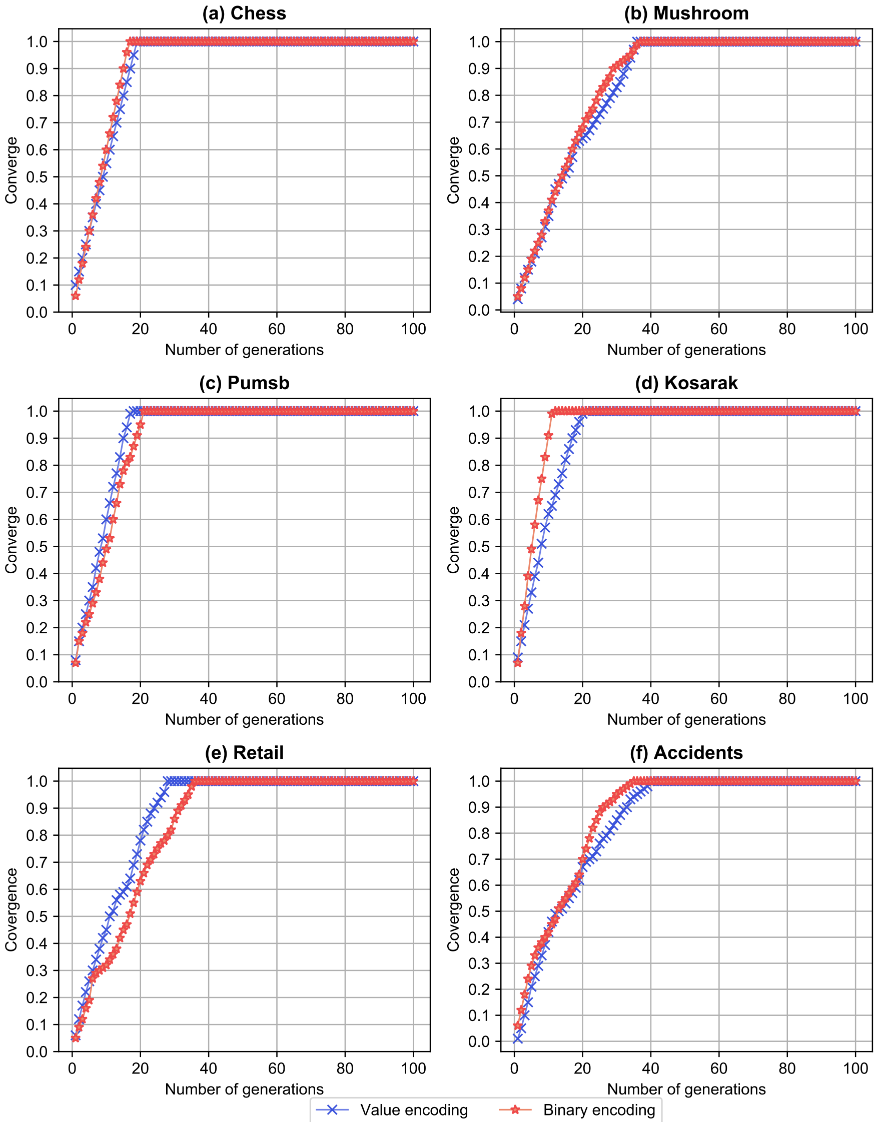


Fig. 2: The Cov of two encoding schemas

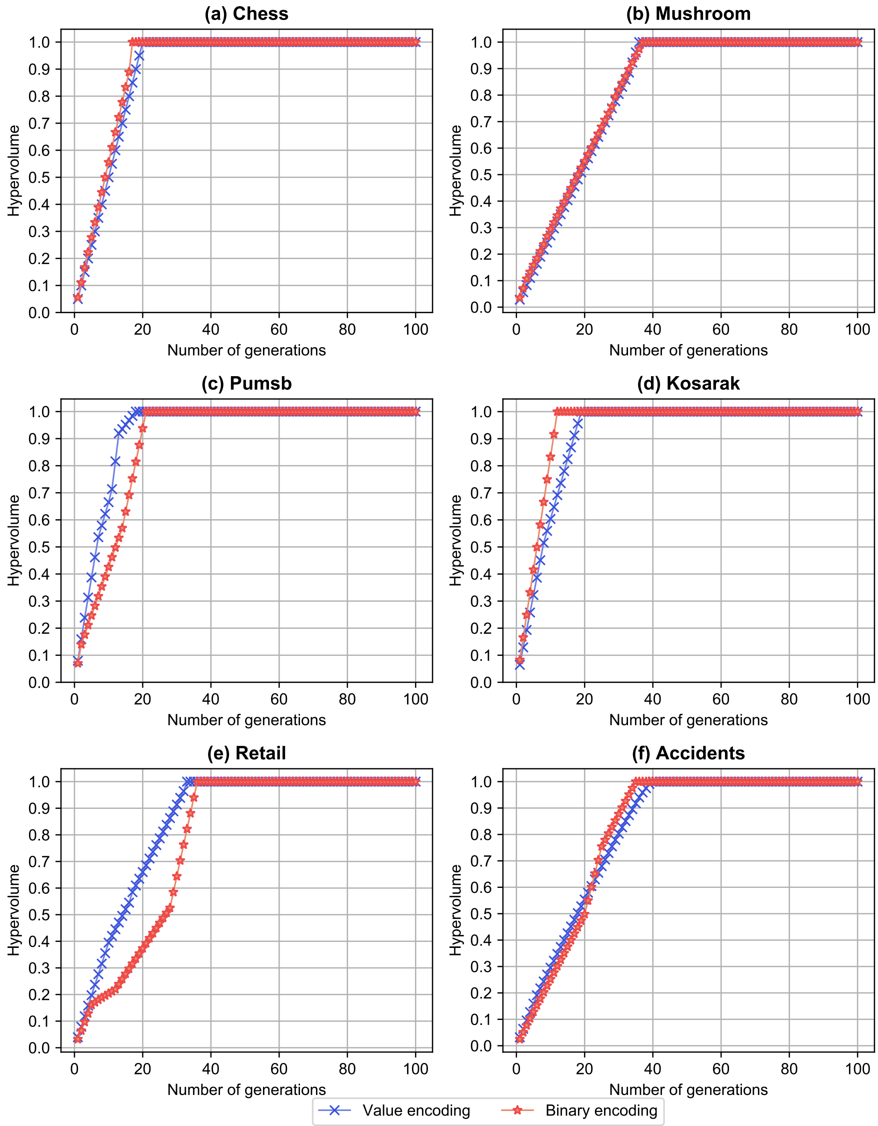


Fig. 3: The HV of two encoding schemas

5.2 Pattern analysis

Then, the conventional basic algorithms U-Apriori [40] and EFIM [41] are compared with the developed model in terms of HV, Cov and the total number of patterns generated. This is done because there is currently no method that accounts for both utility factors and uncertainty elements using evolutionary progress. However, both methods require a priori knowledge to correctly calculate the threshold for either utility or uncertainty. For this reason, we set the

Table 6: The compared algorithms in terms of Cov, HV, and number of patterns

Algorithm	Thresolds	Metrics	Datasets					
			Chess	Mushrooms	Korasak	Pumsb	Accidents	Retails
U-Apriori (minUncertanity)	0.2	*Cov*	0.32	0.24	0.63	0.35	0.39	0.45
		HV	2267.90	5800.12	430.06	3410.52	23996.42	375.19
		# patterns	18562.00	12926.00	26857.00	35035.00	6080.00	17054.00
	0.4	*Cov*	0.11	0.07	0.33	0.12	0.15	0.16
		HV	1920.48	4909.05	281.66	2754.43	20337.12	229.31
		# patterns	6076.00	3631.00	14057.00	12223.00	2322.00	6150.00
	0.6	*Cov*	0.03	0.02	0.14	0.03	0.04	0.05
		HV	1573.19	3732.24	230.45	2257.76	16675.81	187.61
		# patterns	1469.00	875.00	5800.00	3448.00	602.00	1912.00
	0.8	*Cov*	0.00	0.00	0.04	0.01	0.00	0.01
		HV	1165.46	1979.21	179.25	953.30	13006.52	10.97
		# patterns	210.00	163.00	1623.00	555.00	76.00	376.00
MOEA-HEUPM	Binary	*Cov*	1.00	0.80	0.50	0.60	0.60	1.00
		HV	215088.60	94314.83	138219.34	83481.53	83481.53	2117792.17
		# patterns	2900.00	1939.00	509.00	900.00	900.00	10766.71
	Value	*Cov*	1.00	0.65	0.50	1.00	1.00	1.00
		HV	138219.34	146600.20	146600.20	134567.41	134567.41	756963.07
		# patterns	486.00	2039.00	1929.00	500.00	500.00	9818.60
EFM (minUtil)	0.20	*Cov*	0.50	0.50	0.12	0.50	0.50	0.50
		HV	-	-	-	-	-	-
		# patterns	95381.00	4801.00	8.00	34270.00	3756.00	35288968.00

thresholds for the U-Apriori algorithm to range from 20 percent to 80 percent in increments of 20 percent and used a threshold of 20 percent for the EFIM algorithm. This is due to the fact that the other thresholds (i.e., 40 percent, 60 percent, and 80 percent) result in empty patterns (HUIs) being generated during the mining process. In addition, the total number of uncovered patterns is listed in Table 6.

The number of patterns produced by MOEA-HEUPM is five to ten times less than the number of patterns produced by U-Apriori and EFIM combined. This gap is at least five to ten times larger. For example, the number of patterns produced by the proposed technique on six different datasets without any kind of cutoff value varies from 486 to 2,900, as you can see in Table 6, which you can view here. This table is available for your inspection. Nevertheless, the production of U-Apriori fluctuated from a low of 12,926 to a high of 35,035, while the production of EFIM ranged from a low of 67 to a high of 190,573. As a result, U-Apriori and EFIM take more time to compute, while the proposed model converges quickly and provides fewer choice patterns. One possible explanation for this discrepancy is that the model in question has fewer different patterns to choose from.

As you can see in Table 6, the size of the U-Aprior for both measures, i.e., Cov and HV in the dense and sparse data sets, is quite large when we choose a low value for the uncertainty. This is true despite the fact that the sparse dataset contains fewer observations than the dense dataset. This is the case whether the datasets are rich or limited. As the threshold for uncertainty is increased, the magnitude of the U-priority in the data sets decreases. This leads to a significant increase in the number of patterns that need to be constructed, as can be seen in Table 6. For this reason, it is not an efficient method for mining patterns that have a high degree of uncertainty. However, EFIM is not able to obtain the answer when the threshold for usefulness is set higher, which is defined as more than 20 percent of the time. The performance of the algorithm is directly improved by using the Kosarak dataset. In contrast, as you can see in Table 6, the EFIM algorithm only considered 67 possible candidates. When applied to a large number of datasets, the EFIM technique uncovered and identified a significant number of recurring patterns. The reason for this is that not only the aspect of usefulness is considered, but also the aspect of uncertainty. For this reason, neither of the two commonly used approaches is able to deal with the components of uncertainty and utility, which must be combined to obtain meaningful and helpful patterns.

Moreover, the constructed MOEA-HEUPM reaches the highest possible Cov value, which can range from a maximum of 1 to a minimum of 0.50, as shown in Table 6. With respect to HV, the considered model reaches the highest possible value of 138.219 and the lowest possible value of 83.481, respectively. However, compared to the other two paths, the number of patterns formed by this approach is much lower and the derived patterns are the non-dominant solutions that can be used for effective judgments. In the studies, the values of uncertainty and utility in dense databases (such as chess, mushrooms, cosarak, retail, and accidents) tend to be high, resulting in a high value of HV. This is because dense databases contain a lot of information. It is impossible for the records in a dense dataset to overlap because the datasets are so small. A direct consequence of this is that Cov will grow. Both the meta-itemset population initialization strategy and the transaction itemset population initialization method contribute to the higher Cov value by allowing a greater variety of solutions to be obtained after each generation. This allows the value to be increased. As a result, MOEA-HEUPM was able to identify the patterns that had the greatest expected utility by basing its analysis on the size of the Cov and HV variables. Therefore, a meta-item set with two or more bits must be considered to obtain better answers for the sparse data set.

5.3 Scalability analysis

In this section, we will undertake an examination of the scalability of the proposed algorithm in comparison to the other algorithms that have been applied to the synthetic data set $T10I4N4KDXK$ (X is the size of the dataset represented in the number of transactions). The IBM Quest Synthetic Data Generator is the one that is in charge of the creation of the synthetic dataset [43]. Table 7 displays the findings at a variety of predetermined cutoffs. Both the U-Apriori and EFM algorithms generated a far greater number of patterns than the proposed MOEA-HEUPM; nevertheless, the MOEA-HEUPM algorithm had higher performance in terms of Cov and HV. The MOEA-HEUPM algorithm was constructed in comparison to the MOEA-HEUPM that was created. When the threshold for doubtfulness is raised to a higher value, there is a corresponding reduction in the HV as well as the number of patterns. The findings of the EFM analysis show the same trend as well. It is not feasible to estimate the hypervolume for this particular collection of research since EFIM has such a vast number of patterns. In addition, it is necessary to include both utility and uncertainty, which makes the calculation challenging owing

Table 7: Scalability analysis

Algorithm	Thresolds	Metrics	T40I10DXK				
			X=20	*X=40*	*X=60*	*X=80*	*X=100*
U-Apriori (minUncertanity)	*0.2*	Cov	1	0.56	0.39	0.39	0.38
		HV	41818.9	41820.0	41819.5	41819.8	41820.0
		# patterns	164589	380353	624123	624760	625398
	0.4	Cov	1	0.56	0.39	0.39	0.39
		HV	203.9	203.9	203.9	203.9	203.9
		# patterns	30756	71415	117669	118077	118485
	0.6	Cov	1	0.56	0.39	0.38	0.38
		HV	163.0	163.04	163.04	163.06	163.03
		# patterns	8197	18808	30840	31080	31318
	0.8	Cov	1	0.51	0.34	0.28	0.24
		HV	120.69	120.04	121.50	121.33	121.89
		# patterns	116	239	365	446	523
MOEA-HEUPM	Binary	*Cov*	0.97	0.97	0.96	0.97	0.98
		HV	2013667.9	4110374.0	4624663.4	8113592.7	7249258.8
		# patterns	4182	3570	2754	3774	6222
	Value	*Cov*	0.97	0.96	0.97	0.97	0.97
		HV	1515809.4	4109259.5	4622814.3	6111761.8	9586723.6
		# patterns	4998	3264	4386	4284	4284
EFM (minUtility)	0.8	*Cov*	1	0.74	0.69	1	0.69
		HV	-	-	-	-	-
		# patterns	71705	250558	733474	239405	518260

to the high number of transactions and the enormous variety of patterns that may be retrieved from EFIM. The threshold in EFIM is set rather high since it is the component that is responsible for creating a significant amount of pattern data. The U-Apriori and EFIM models provide much greater convergence rates and a bigger number of patterns when the total number of transactions in the synthetic dataset is raised to $20,000$. On the other hand, the Cov falls to a lower value as the number of transactions rises, which eventually results in more savings. Despite this, the MOEA-HEUPM that has been suggested still achieves acceptable levels of performance, in spite of the growing size of the database. According to the findings of the empirical research, it seems that the model that was provided is useful in removing the requirement for a threshold in order to extract a set of uncertain item sets that have high utility.

6 Conclusion

Mining by relevant patterns, as opposed to the more typical mining by association rules that has attracted much attention in recent decades, could lead to finding more useful information. This is one of the advantages of data mining. Due to the rapid technological progress, data uncertainty is recognized as an important factor in the field of pattern mining. However, most of the existing and general solutions require setting a threshold to identify the needed information. This is not a simple process and is considered inappropriate in a variety of domains and applications. In the first part of this paper, we consider the utility and uncertainty components together. Then, based on MOEA/D, we construct an evolutionary model, which we call MOEA-HEUPM, to find the non-dominant patterns that have high expected utility (HEUPs). Finally, we will discuss the results of our work. To show the utility of the newly created MOEA-HEUPM, we use two different coding schemes: a binary one and a value-based one. To make decisions in an uncertain environment, avoiding the use of a priori thresholds, it is more effective to construct fewer but more helpful non-dominated HEUPs based on the created MOEA-HEUPM. Experiments are then conducted to illustrate the effectiveness of the proposed model compared to the general and conventional U-Apriori and EFIM models in terms of convergence, hypervolume, and the number of patterns that can be detected.

7 Future Work

Since the multi-objective model presented is the first work to consider both objectives, utility and uncertainty, in a database that contains uncertainty, a number of new possibilities arise that can be explored in further research. For example, the model is able to account for tuple unpredictability with respect to the multiple objectives problem. This is possible because each tuple is provided with its own probability distribution. When solving multiple objective problems, it is sometimes feasible to consider a larger number of features to produce a smaller number of suitable non-dominant patterns. Another intriguing topic that could be explored is the possibility of extending the MOEA-based model just presented to the areas of dynamic data mining, stream data mining, and top-k pattern mining.

References

[1] Fournier-Viger, P. *et al.* A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **7** (4), e1207 (2017). https://doi.org/10.1002/widm.1207 .

[2] Agrawal, R. & Srikant, R. *Fast algorithms for mining association rules in large databases*, 487–499 (1994).

[3] Fournier-Viger, P., Wu, C., Zida, S. & Tseng, V. S. *FHM: faster high-utility itemset mining using estimated utility co-occurrence pruning*, Vol. 8502, 83–92 (2014).

[4] Gan, W. *et al.* A survey of utility-oriented pattern mining. *IEEE Transactions on Knowledge and Data Engineering* 1–1 (2019). https://arxiv.org/abs/1805.10511 .

[5] Gan, W., Lin, J. C. W., Fournier-Viger, P., Chao, H. & Yu, P. S. A survey of parallel sequential pattern mining. *ACM Transactions on Knowledge Discovery from Data* **13** (3), 25–34 (2019). https://doi.org/10.1145/3314107 .

[6] Liu, Y., keng Liao, W. & Choudhary, A. *A fast high utility itemsets mining algorithm*, 90–99 (2005).

[7] Liu, M. & Qu, J. *Mining high utility itemsets without candidate generation*, 55–64 (2012).

[8] Ryang, H. & Yun, U. Top-k high utility pattern mining with effective threshold raising strategies. *Knowledge-Based Systems* **76**, 109–126 (2015). https://doi.org/10.1016/j.knosys.2014.12.010 .

[9] Kannimuthu, S. & Premalatha, K. Discovery of high utility itemsets using genetic algorithm with ranked mutation. *Applied Artificial Intelligence* **28** (4), 337–359 (2014). https://doi.org/10.1080/08839514.2014.891839 .

[10] Lin, J. C. W. *et al.* Mining high-utility itemsets based on particle swarm optimization. *Engineering Applications of Artificial Intelligence* **55**, 320–330 (2016). https://doi.org/10.1016/j.engappai.2016.07.006 .

[11] Lin, J. C. W., Yang, L., Fournier-Viger, P., Hong, T. & Voznák, M. A binary PSO approach to mine high-utility itemsets. *Soft Computing* **21** (17), 5103–5121 (2017). https://doi.org/10.1007/s00500-016-2106-1 .

[12] Wu, J. M. T., Zhan, J. & Lin, J. C. W. An ACO-based approach to mine high-utility itemsets. *Knowledge-Based Systems* **116**, 102–113 (2017). https://doi.org/10.1016/j.knosys.2016.10.027 .

[13] Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S. & Coello, C. A. C. Survey of multiobjective evolutionary algorithms for data mining: Part II. *IEEE Transactions on Evolutionary Computation* **18** (1), 20–35 (2014). https://doi.org/10.1109/TEVC.2013.2290082 .

[14] Ryerkerk, M., Averill, R. C., Deb, K. & Goodman, E. D. A survey of evolutionary algorithms using metameric representations. *Genetic Programming and Evolvable Machines* **20** (4), 441–478 (2019). https://doi.org/10.1007/s10710-019-09356-2 .

[15] Zou, F., Chen, D., Huang, D. S., Lu, R. & Wang, X. Inverse modelling-based multi-objective evolutionary algorithm with decomposition for community detection in complex networks. *Physica A: Statistical Mechanics and its Applications* **513**, 662–674 (2019). https://doi.org/10.1016/j.physa.2018.08.077 .

[16] Wang, Z., Zhang, Q., Zhou, A., Gong, M. & Jiao, L. Adaptive replacement strategies for MOEA/D. *IEEE Transactions on Cybernetics* **46** (2), 474–486 (2016). https://doi.org/10.1109/TCYB.2015.2403849 .

[17] Deb, K., Agrawal, S., Pratap, A. & Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **6** (2), 182–197 (2002). https://doi.org/10.1109/4235.996017 .

[18] Shi, R. & Lee, K. Y. Multi-objective optimization of electric vehicle fast charging stations with SPEA-II. *IFAC-PapersOnLine* **48** (30), 535–540

(2015). https://doi.org/10.1016/j.ifacol.2015.12.435 .

[19] Zhang, L., Fu, G., Cheng, F., Qiu, J. & Su, Y. A multi-objective evolutionary approach for mining frequent and high utility itemsets. *Applied Soft Computing* **62**, 974–986 (2018). https://doi.org/10.1016/j.asoc.2017.09.033 .

[20] Djenouri, Y., Fournier-Viger, P., Belhadi, A. & Lin, J. C. in *Metaheuristics for frequent and high-utility itemset mining* 261–278 (Springer, 2019).

[21] Lin, J. C. W., Gan, W., Fournier-Viger, P., Hong, T. & Tseng, V. S. Efficient algorithms for mining high-utility itemsets in uncertain databases. *Knowledge-Based Systems* **96**, 171–187 (2016). https://doi.org/10.1016/j.knosys.2015.12.019 .

[22] Lee, G. & Yun, U. A new efficient approach for mining uncertain frequent patterns using minimum data structure without false positives. *Future Generation Computer Systems* **68**, 89–110 (2017). https://doi.org/10.1016/j.future.2016.09.007 .

[23] Lee, G., Yun, U. & Ryang, H. An uncertainty-based approach: Frequent itemset mining from uncertain data with different item importance. *Knowledge-Based Systems* **90**, 239–256 (2015). https://doi.org/10.1016/j.knosys.2015.08.018 .

[24] Liu, Y. Mining time-interval univariate uncertain sequential patterns. *Data & Knowledge Engineering* **100**, 54–77 (2015). https://doi.org/10.1016/j.datak.2015.07.012 .

[25] Palacios, A. M., Martínez, A., Sánchez, L. & Couso, I. Sequential pattern mining applied to aeroengine condition monitoring with uncertain health data. *Engineering Applications of Artificial Intelligence* **44**, 10–24 (2015). https://doi.org/10.1016/j.engappai.2015.05.003 .

[26] Lin, J. C. W., Gan, W., Fournier-Viger, P., Hong, T. & Tseng, V. S. Weighted frequent itemset mining over uncertain databases. *Applied Intelligence* **44** (1), 232–250 (2016). https://doi.org/10.1007/s10489-015-0703-9 .

[27] Ahmed, A. U., Ahmed, C. F., Samiullah, M., Adnan, N. & Leung, C. K. S. Mining interesting patterns from uncertain databases. *Information Sciences* **354**, 60–85 (2016). https://doi.org/10.1016/j.ins.2016.03.007 .

[28] Bui, N., Vo, B., Huynh, V., Lin, C. W. & Nguyen, L. T. T. *Mining closed high utility itemsets in uncertain databases*, 7–14 (2016).

[29] Yun, U., Nam, H., Lee, G. & Yoon, E. Efficient approach for incremental high utility pattern mining with indexed list structure. *Future Generation Computer Systems* **95**, 221–239 (2019). https://doi.org/10.1016/j.future.2018.12.029 .

[30] Yun, U., Lee, G. & Yoon, E. Efficient high utility pattern mining for establishing manufacturing plans with sliding window control. *IEEE Transactions on Industrial Electronics* **64** (9), 7239–7249 (2017). https://doi.org/10.1109/TIE.2017.2682782 .

[31] Zhang, L., Luo, P., Chen, E. & Wang, M. Revisiting bound estimation of pattern measures: A generic framework. *Information Sciences* **339**, 254–273 (2016). https://doi.org/10.1016/j.ins.2015.12.036 .

[32] Mai, T., Nguyen, L. T. T., Vo, B., Yun, U. & Hong, T. Efficient algorithm for mining non-redundant high-utility association rules. *Sensors* **20** (4), 1078 (2020). https://doi.org/10.3390/s20041078 .

[33] Baek, Y., Yun, U., Yoon, E. & Fournier-Viger, P. Uncertainty based pattern mining for maximizing profit of manufacturing plants with list structure. *IEEE Transactions on Industrial Electronics* 1–1 (2019). https://doi.org/10.1109/tie.2019.2956387 .

[34] Lee, G., Yun, U. & Lee, K. Analysis of tree-based uncertain frequent pattern mining techniques without pattern losses. *The Journal of Supercomputing* **72** (11), 4296–4318 (2016). https://doi.org/10.1007/s11227-016-1847-z .

[35] Lin, J. C. W., Gan, W., Fournier-Viger, P., Hong, T. & Tseng, V. S. Efficiently mining uncertain high-utility itemsets. *Soft Computing* **21** (11), 2801–2820 (2017). https://doi.org/10.1007/s00500-016-2159-1 .

[36] Gan, W., Lin, J. C. W., Fournier-Viger, P., Chao, H. C. & Tseng, V. S. *Mining high-utility itemsets with both positive and negative unit profits from uncertain databases*, Vol. 10234 of *Lecture Notes in Computer Science*, 434–446 (2017).

[37] Kumar, A. Encoding schemes in genetic algorithm. *The International Journal of Advanced Research in IT and Engineering* **2** (3), 1–7 (2013) .

[38] Mohammadi, F. G., Amini, M. H. & Arabnia, H. R. in *Evolutionary computation, optimization, and learning algorithms for data science* 37–65 (2020).

[39] Zhang, X., Tian, Y., Cheng, R. & Jin, Y. An efficient approach to non-dominated sorting for evolutionary multiobjective optimization. *IEEE Transactions on Evolutionary Computation* **19** (2), 201–213 (2015). https:

//doi.org/10.1109/TEVC.2014.2308305 .

[40] Hooshsadata, M., Bayat, S., Naeimi, P., Mirian, M. S. & Zaiane, O. R. Uapriori: an algorithm for finding sequential patterns in probabilistic data. *Uncertainty Modeling in Knowledge Engineering and Decision Making* 907–912 (2012). https://doi.org/10.1142/9789814417747_0145 .

[41] Zida, S., Fournier-Viger, P., Lin, J. C. W., Wu, C. & Tseng, V. S. *EFIM: A highly efficient algorithm for high-utility itemset mining*, Vol. 9413, 530–546 (2015).

[42] Fournier-Viger, P. *et al.* *The SPMF open-source data mining library version 2*, Vol. 9853, 36–40 (2016).

[43] Agrawal, R. *et al.* *The quest data mining system*, 244–249 (1996).