Dhinaharan Nagamalai
Natarajan Meghanathan (Eds)

Computer Science & Information Technology

7th International Conference on Information Technology Convergence and Services (ITCSE 2018) May 26~27, 2018, Vienna, Austria.

AIRCC Publishing Corporation

Volume Editors

Dhinaharan Nagamalai,
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

ISSN: 2231 - 5403
ISBN: 978-1-921987-85-4
DOI : 10.5121/csit.2018.80701 - 10.5121/csit.2018.80707

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The 7th International Conference on Information Technology Convergence and Services (ITCSE 2018) was held in Vienna, Austria during May 26~27, 2018. The 10th International Conference on Wireless & Mobile Network (WiMo 2018), The 7th International Conference on Advanced Computer Science and Information Technology (ICAIT 2018), The 7th International Conference on Digital Image Processing and Vision (ICDIPV 2018) and The 7th International Conference on Cryptography and Information Security (CRYPIS 2018) was collocated with The 7th International Conference on Information Technology Convergence and Services (ITCSE 2018). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ITCSE-2018, WiMo-2018, ICAIT-2018, ICDIPV-2018, CRYPIS-2018 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, ITCSE-2018, WiMo-2018, ICAIT-2018, ICDIPV-2018, CRYPIS-2018 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the ITCSE-2018, WiMo-2018, ICAIT-2018, ICDIPV-2018, CRYPIS-2018

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Dhinaharan Nagamalai
Natarajan Meghanathan

Organization

General Chair

David C. Wyld	Southeastern Louisisna University, USA
Jan Zizka	Mendel University in Brno, Czech Republic

Program Committee Members

Abdelhalim Boutarfa	University of Batna, Algeria
Abderrahmane Nitaj	University of Caen Normandie, France
Annamalai	Prairie View A&M University, USA
Artyom Grigoryan	University of Texas at San Antonio, USA
Asmaa Shaker Ashoor	Babylon University, Iraq
Azam Khalili	Malayer University, Iran
Babar Shah	Zayed University ,UAE
Belal M. Abuata	Yarmouk University, Jordan
Bhuvan Modi	Senior Member of Technical Staff, AT&T, USA
Boukenadil Bahidja	University of Tlemcen, Algeria
Cristina-Loredana Duta	University Politehnica of Bucharest, Romania
Dimitris Kontoudis	University of Macedonia, Greece
El Miloud Ar-reyouchi	Abdelmalek Essaadi University and SNRT, Morocco
Emad Eldin Mohamed	Canadian University Dubai, UAE
Emilio Jiménez Macías	University of La Rioja, Spain
Ewerton Romulo Silva Castro	Instituto Federal da Paraiba (IFPB), Brazil
Fatma Outay	Zayed University DXB, UAE
Firdous Imam	Saud Islamic University, Saudi Arabia
Gharbaoui Molka	Scuola Superiore Sant'Anna, Italy
Ghassan Qas marrogy	Cihan University, Iraq
Hamid Ali Abed AL-Asadi	Basra University, Iraq
Hamzeh Khalili	Universitat Politecnica de Catalunya (UPC), Spain
Hani Bani-Salameh	The Hashemite University, Jordan
Hari Krishna Garg	National University of Singapore, Singapore
Hongzhi Harbin	Institute of Technology, China
Husnain Sherazi	Polytechnic University of Bari, Italy
Isa Maleki	Islamic Azad University, Iran
Jamel Hattay	University of Tunis, Tunisis
Joberto S. B. Martins	Salvador University, Argentina
Khalid mohamed oqlah nahar	Yarmouk University, Jordan
Kishan Rao Kalitkar	Vaagdevi Group of Technical Institutions, India
Liyakathunisa Syed	Prince Sultan University, Saudi Arabia
Mike Turi	California State University-Fullerton, USA
Miroslaw Kwiatkowski	AGH University of Science and Technology, Poland
Mohammad Zarour	Prince Sultan University, Kingdom of Saudi Arabia
Mohammed Al-Sarem	Taibah University, KSA
Mohammed Nabil EL KORSO	Paris Nanterre University, France

Mohanned Akour	Yarmouk University, Jordan
Mokhtar Mohammadi	Shahrood University of Technology, Iran
Muddu Madakyaru	Manipal University, India
Muneer Masadeh Bani Yassein	Jordan University of Science and Technology, Jordan
Nadhir Ben Halima	Taibah University, Saudi Arabia
Nahlah M. Ameen Shatnawi	Yarmouk University, Jordan
Nahlah Shatnawi	Yarmouk University, Jordan
Nayan M. Kakoty	Tezpur University, India
Nayeem Ahmad Khan	University Malaysia Sarawak, Malaysia
Noura Taleb	Badji Mokhtar University, Algeria
Orhan Dagdeviren	Ege University, Turkey
Ouafa Mah	Ouargla university, Algeria
Paulo Roberto Martins de Andrade	University of Regina, Canada
Pelin Angin	Purdue University, USA
Prakash Duraisamy	University of Central Missouri, United States
Rahil Hosseini	Islamic Azad University, Iran
Rajoua Anane Bouallegue	Université du Maine,France
Ramgopal Kashyap	Sagar Institute of Science and Technology, India
Rana Rahim	Lebanese University, Lebanon
Razieh malekhoseini	Islamic Azad University, Iran
Rhattoy	Moulay Ismail University, Morocco
Riccardo Pecori	eCampus University, Italy
Roberto Paiano	University of Salento, Italy
Saban Gulcu	Necmettin Erbakan University, Turkey
Sabina Rossi	University in Venice, Italy
Saed Tarapiah	An-najah National University, Palestine
Salem Hasnaoui	National Engineering School of Tunisi, Tunisia
Saltanat Meiramova	Seifullin Kazakh Agrotechnical University, Kazakhstan
Samir Bandyopadhyay	University of Calcutta, India
Samuel Baraldi Mafra	Federal University of Parana (UFPR), Brazil
Sandra D'Souza	Manipal University, India
Shadi R. Masadeh	Isra University, Jordan
Shameem SSS	Manipal International University, Malaysia
Sos Agaian	University of Texas at San Antonio, USA
Stefano Michieletto	University of Padova, Italy
Taeghyun Kang	University of Central Missouri, United States
Tanzila Saba	Prince Sultan University, Kingdom of Saudi Arabia
Tayebeh Askari	Bam University of Iran, Iran
Tonghan Wang	East China Institute of Technology, China
Tribikram Pradhanaaa	Manipal Institute of Technology, India
Upasna Vishnoi	Marvell Technology Group Ltd, USA
Victor Banos	Technical University of Catalonia, Spain
W'ael Alzyadat	Isra University, Jordan
Wail Mardini	Jordan University of Science and Technology, Jordan
Xuechao Li	Auburn University, USA
Yao-Nan Lien	Asia University, Taiwan
Zeyu Sun	Luoyang Institute of Science and Technology, China
Zsolt Polgar	Technical University of Cluj Napoca, Romania

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Networks & Communications Community (NCC)

Soft Computing Community (SCC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

7th International Conference on Information Technology Convergence and Services (ITCSE 2018)

10th International Conference on Wireless & Mobile Network (WiMo 2018)

7th International Conference on Advanced Computer Science and Information Technology (ICAIT 2018)

7th International Conference on Digital Image Processing and Vision (ICDIPV 2018)

7th International Conference on Cryptography and Information Security (CRYPIS 2018)

AN INTEROPERABLE CLOUD-BASED GEOPORTAL FOR DISCOVERY AND MANAGEMENT OF EARTH OBSERVATION PRODUCTS

Dante D. Sánchez-Gallegos[1], J.L. Gonzalez-Compean[1], Victor J. Sosa-Sosa[1], Heidy M. Marin-Castro[2] and José Tuxpan-Vargas[3]

[1]Cinvestav Tamaulipas, Cd. Victoria, Tamaulipas, México
[2]Cátedras CONACYT -UAT, Cd. Victoria, Tamaulipas, México
[3]IPICyT, San Luis Potosí, SLP, México

ABSTRACT

This paper presents the design and development of an interoperable geoportal service for discovery and management of earth observation products (EOPs). In this service, the geoportal components are encapsulated into virtual containers that are launched into the cloud by using a microservice scheme to solve issues such as interoperability (with other systems) and implementation (over different platforms). A search microservice that analyses the preferences of end-users (settings of spatiotemporal and polygon shapes) and builds clusters of users sharing preferences was included into the geoportal for recommending/delivering, in advance, products matching with end-user preferences. The geoportal service also enables end-users to organize EOPs on-the-fly by using spatiotemporal parameters. A prototype of this service was implemented in a private cloud and connected to a satellite imagery repository of an antenna (ERIS) managed by Mexican Space Agency in a proof of concept. Learned lessons and performance assessments are described through an experimental evaluation with real users' participation.

KEYWORDS

Spatial data management, Satellite imagery repositories, User Preferences, Data Analysis & Geoportals

1. INTRODUCTION

Geoportals are key tools for space agencies to show and delivery earth observation products (EOPs) to end-users. Catalogs of satellite images, derivative products (corrected and thematic products) and maps [1], as well as maps created on-the-fly by end-users [2] are examples of EOPs. These products represent assets for agencies, scientific community and government instances to conduct missions, research and programs about earth respectively. EOPs are used in studies about climatic and disaster event management, territory management strategies as well as environment studies.

Dhinaharan Nagamalai et al. (Eds) : ITCSE, WiMo, ICAIT, ICDIPV, CRYPIS - 2018
pp. 01–13, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80701

However, the development and deployment of these tools on the cloud/web when managing large satellite imagery repositories is not trivial because of technical and management issues.

The technical challenges are related to the interoperability of the geoportals with other systems. A geoportal is the last stage in a life cycle of satellite imaginary where additional software is also used depending on the stages of that cycle managed by agencies. In these stages EOPs are transformed into new versions (derivative products, maps, etc) or new products are created by grouping EOPs using spatial and temporal parameters. For instance, in the acquisition stage, services are required for parsing metadata files created by antenna operators, mapping with EOPs and make data and metadata indexes available for other services/applications and geoportals. In the preservation [3] stage indexed EOPs are ensured/stored in conservation infrastructure (I.e. Cloud/cluster storage services) for other systems to acquire these products. In manufacturing [4] stage the preserved products are used to create new EOPs that are indexed by geoportals and made available to other services.

These services, systems and applications are commonly created as either a monolithic application or independent solutions using different methods of access and data exchange. In the first case, the building of solutions passing through different life cycle stages is not trivial, whereas in the second case, designers must adapt their geoportal to each system with which it exchanges data/metadata. Moreover, these systems are commonly developed as web services deployed mainly on the cloud side by using virtual machines that, depending on the size of the geoportal, could become unmanageable [5].

In turn, the virtual container represents an alternative to virtual machines for solving issues of deployment on different types of infrastructures in an efficient manner [5]. When these containers are added to microservices including standardized input/output data exchange, interoperability issues are solved as well as different solutions including different stages of life cycle can be built. However, to the best of our knowledge, architectures of geoportals based on microservices and containers have not been enough explored.

Besides the interoperability and flexibility issues, the information management remains as a challenge for space agencies. The number and type of EOPs available in geoportals depend on the transformation depend on the transformations applied to EOPs, required by an agency, during the stages of the life cycle of the satellite imagery. In this context, product discovery schemes could enable geoportals to exhibit useful products and to improve the service experience of the end-users. Traditionally, geoportals include information searchers based mainly on a spatial criterion defined either by polygon shapes and/or parameters (e.g. Path and row, decimals or coordinates) [6]. This type of search management is mainly suitable for researchers and academics. Although recently geoportals are also incorporating searchers using natural language [6], there still is an opportunity area to improve the experience service of end-users by analysing preferences of end-users and their relationships with the preferences of other researchers/academics.

This paper presents the design and development of an interoperable cloud-based geoportal for the discovery and management of earth observation products. A microservice scheme was developed to encapsulate the applications used by geoportals in the life cycle of EOPs into virtual containers, which are launched into the cloud to create a single unified service of chained microservices.

This scheme enables agencies to create processing workflows as well as to solve interoperability and deployment issues. A search engine was created as a microservice connected to the geoportal to analyse the preferences of end-users (settings of spatiotemporal and polygon shapes) and to build clusters of users sharing preferences. This microservice was included into a geoportal for recommending/delivering, in advance, products matching with end-user preferences. The geoportal also enables end-users to create, on-the fly, products by using spatiotemporal parameters, which are also indexed by the geoportal and the search engine.

A prototype of the geoportal service was implemented in a private cloud and connected to a satellite imagery repository service of an antenna (ERIS) managed by Mexican Space Agency by using a federated service [3, 7]. A proof of concept was performed with this prototype and an experimental evaluation was conducted with the participation of real end-users. The evaluation revealed the feasibility of developing interoperable cloud-based geoportals by using microservices and containers. Learned lessons and performance assessments about management and discovery of EOPs are also described.

This paper is organized as follows: Section 2 introduces the design principles, including development details using microservices and provides an explanation of the method for detecting user preference; Section 3 and 4 present experimental evaluation and results respectively, and finally Section 5 provides some conclusions and future work.

2. AN INTEROPERABLE CLOUD-BASED GEOPORTAL: DESIGN PRINCIPLES

In this section, we present the general design of an interoperable geoportal service designed for the Mexican Space Agency (AEM), to exhibit EOPs through the cloud.

Figure 1 shows a conceptual representation of the containers of a geoportal implemented by using a microservice scheme. As it can be seen, the microservice structure includes three virtual containers (front-end, service and back-end) chained in a pipe and filter pattern. The front-end is in charge of receiving data or metadata from applications/services or other microservice and makes indirection of this information to service containers in a load balancing manner. The front-end orchestrates the launching and coordination of service and back-end containers in a workflow. The services container performs the choreography sent by front-end, which enables administrators to create software patterns. The service container of the geoportal includes three micro-services (Access, management and storage) that are chained building a stack structure defined by the orchestration configuration in Front-end container. Other patterns such as pipe and filter or master and slave can be deployed in the service container.

The containers in the service stack processes the requests sent through front-end and delivers the results through back-end container, which was designed for receiving and/or delivering data and/or metadata to/from either a client application or other microservice. In the case shown in Figure 1, geoportal service is connected to the service for the preservation of a satellite imagery, which is used to preserve the EOPs captured by the ERIS antenna.

In this scheme, each container also includes its dependencies and even a reduced operating system to ensure a regular operation over different types of platforms; as a result, the geoportal can be replicated to build clusters to improve the experience of end-users [8]. Moreover, the

service container can be reused to change the geoportal by another application (e.g. preservation of images or manufacturing of derivative products) and could enable agencies to create workflows through the life cycle of the satellite imagery [3].

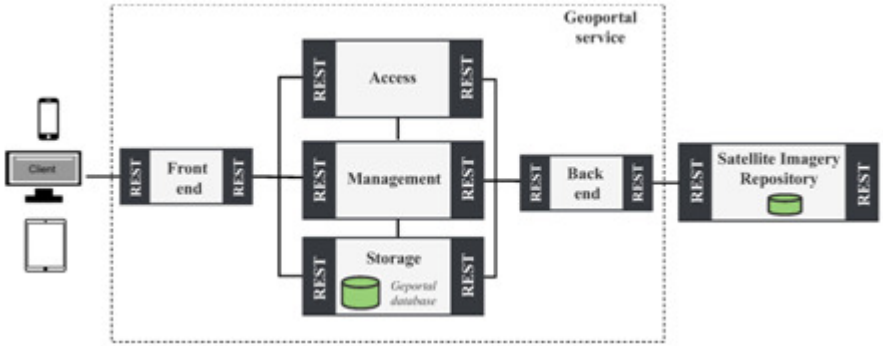

Figure 1: Conceptual representation of the containers of a geoportal implemented by using a microservice scheme

2.1. Geoportal microservices: development details

The development details of access, management, and storage microservices designed for Geoportal service are described in this section.

The access microservice allows geoportal to manage requests sent by either end-users or apps to the geoportal service through a web browser or invoking them by a REST API. An access control module provides end-users/applications with valid encrypted credentials and tokens to get access to the management microservice.

The management microservice includes a catalog management based on Pub/Sub system, a search engine, and on-the-fly building of EOP system. The catalog management system enables owners of EOPs to perform operations such as create, update, delete, modify and list catalogs. This system also enables users/APPs to share EOPs, found by the geoportal, with other active users. The search engine and the on-the-fly EOPs builder are described in next sections.

The storage microservice includes a retrieval system that enables to exchange data with repositories and cloud storage services.

2.2. Discovering EOPs by Analysing Activity and Preferences of End-Users

The search engine created for end-users to discover EOPs in the geoportal service includes modules such as recommendation and log activity.

The log activity module registers the actions performed by the end-users. It considers each request sent to the management microservice by applications as well as each click done by the end-users on the items of the geoportal web page.

The recommendation module creates suggestion lists of different versions of EOPs indexed by the geoportal for each user based on the user's activity, which reveals preferences (a type of EOPs, catalogs, land cover, timelines, etc.).

A collaborative filtering technique following the memory-based approach was applied to the recommendation module and it is depicted in Figure 2. A recommendation is produced through phases such as logging historical activity of end-users, inference about the zone of interest, classification of active users, and delivering collaborative recommendations to the end-user through the exhibition web service of the geoportal. A profile for each end-user is created in this procedure, which includes the actions performed by the geoportal users, such as the number of views, clicking and downloading of each EOP, the configurations of each query that the user has done when performing searches as well as the user location. The first time that an end-user access to the geoportal, this automatically starts the construction of an activity profile, which is updated each time that user accesses to the geoportal service.

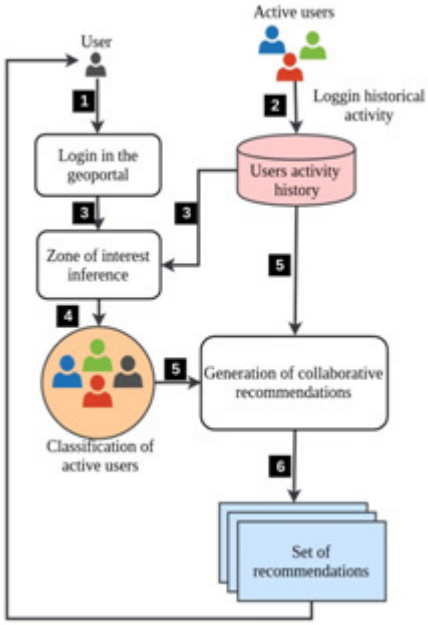

Figure 2: Workflow to create collaborative recommendations

In searches based on polygons, a set of geographic coordinates delimiting a land cover of interest is stored. In the case of using a circle to select an area, the radio and the coordinates defining the center of that area are stored. For points of reference, the latitude and longitude delimiting that area are considered and stored. An inference of the coverage land areas in which the users are interested in is performed as the end-users could be interested in specific coverage areas for long periods of time. In this context, the centroids of the areas where end-users perform queries for the geoportal represent a highly possible zone of interest for the end-user.

2.2.1. Classification of Active Users

A classification of active users is performed as the recommendations for a given user are created by using the information of other end-users exhibiting similar preferences to that users. The centroids of the zones of interest previously calculated are passed to a clustering algorithm called DBSCAN that was included in the recommendation module to create groups of users that share preferences by using the Haversine distance.

When the clusters of users are created, the recommendations for a specific user are built and sent to the geoportal. This task is performed by the Algorithm presented in Figure 3, which receives as input the cluster in which the user was grouped, and the set of EOPs with which the user has made contact (clicked, viewed, downloaded).

Algorithm 1 Collaborative filtering recommender

Require: $N = \{n_1, n_2, n_3, ..., n_n\}$: group (cluster) of neighbor users of user u;
$I = \{i_1, i_2, i_3, ..., i_m\}$: set of current EOP queried by u.
Ensure: $R = \{r_1, r_2, r_3, ..., r_k\}$: set of recommended items (EOP) for u.

```
1:  items = {}
2:  for all n ∈ N do
3:      clicked = n.getClickedEOPs()
4:      downloaded = n.getDownloadedEOPs()
5:      for all EOP ∈ clicked or EOP ∈ downloaded do
6:          if EOP ∉ I then
7:              incrementCounter(EOP.id)
8:              items = items ⋃ EOP
9:          end if
10:     end for
11: end for
12: sort(items)
13: R = selectTop(items)
14: return R
```

Figure 3: Collaborative filtering recommender

The algorithm selects the most popular images previously queried by using the historical activity of the users in the cluster. A table of items is constructed with the EOP identifier (a counter indicating the times that the EOP appears in the group) and the EOP that was clicked or downloaded. Finally, the table is sorted by the number of occurrences of a given EOP, and a configurable number of top records are returned as recommended items.

2.3. Organizing EOPs on-the-fly

The users can organize EOPs on-the-fly, through the geoportal, in the form of mosaics (spatial settings), overlaps (temporal settings), and mosaics overlapped (spatiotemporal settings).

Mosaics are EOPs created for a given land cover selected by an end-user. The images that are within the area, are ordered sequentially according to their coordinates (see Figure 4). The overlaps of images are built with a group of images with the same Path/Row value and sorted by their acquisition date and a timeline for this coverage area is created for end-users to observe changes in coverage land over the time. Mosaics overlapped are created for end-users to observe images when using spatiotemporal parameters.

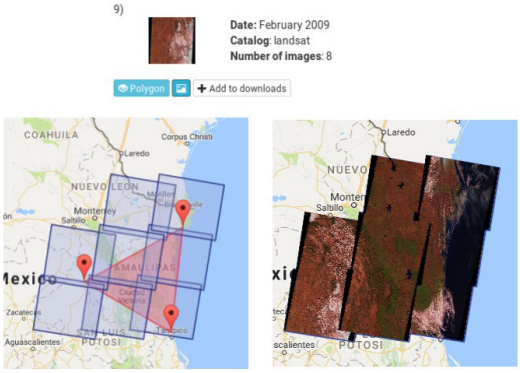

Figure 4: Organizing EOPs in the form of mosaics for a search delimited by an end-user

3. EXPERIMENTAL EVALUATION METHODOLOGY

In this section, we describe the methodology used to conduct a proof of concept and an experimental evaluation based on a satellite imagery of an antenna.

3.1. The Geoportal deployment on a private cloud

The Geoportal service was deployed on a private cloud built by using OpenStack Mitaka. This service was deployed by using a master/slave pattern (See Figure 5) where the microservices of the geoportal were cloned five times and launched into the cloud to build a cluster of five Geoportal slave services and one dispatcher intermediary microservice representing the master of the cluster. The master was in charge of distributing the requests/users to the cluster of slaves in a load balancing manner. The dispatcher also included a centralized database that is accessed by each geoportal microservice, which also includes a front-end container to receive data from the dispatcher and a back-end container to get access to the repository of EOPs and the metadata files as well as to deliver this information to the end-users through the front-end container.

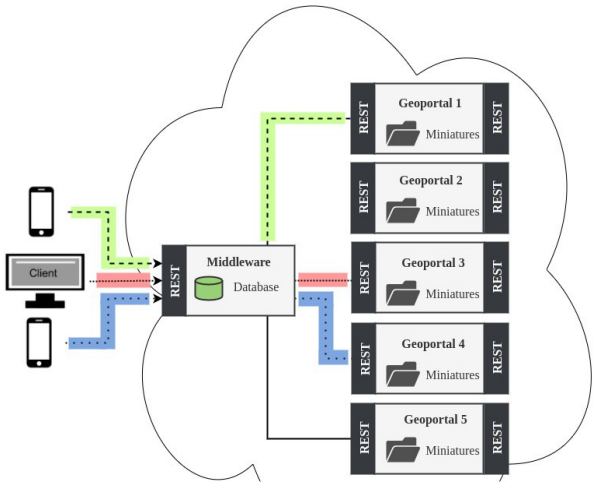

Figure 5: Geoportal distribution

In the storage microservice of the slaves, modules retrieve/delivers EOPs from catalogs preserved and transported by using a federated cloud storage service called FedIDS[3]. The catalogs of satellite images of three sensors, Terra, Aqua of the MODIS platform and LandSat5, were considered in this evaluation.

The containers were built by using Docker. A bot client, was deployed on a virtual machine in the cloud to create workload of end-users that the microservices of the geoportal assumed as a real workload.

3.2. Metrics

The metrics chosen to evaluate the performance of the prototype were:

- *Service time*: Represents the time spent in completing a task by a component of the geoportal, such as master, slave or microservice.

- *Response time*: When performing experiments with real users, this metric represents the time observed by end-users when sending requests to the geoportal. When the bot produces requests in automatic manner, this metric considers the time spent by the bot client to send requests to the dispatcher, the time required by the geoportal slave to serve a given request and the time spent by the dispatcher in spending the response to the bot.

4. EVALUATION RESULTS

In the evaluation performed with client bot, the tests were carried out to measure the performance of the geoportal service when the geoportal slaves produce recommendations for the end-users. The service spent 0.61 seconds to serve a simple query drawing a polygon over some areas of Mexico and 0.93 seconds required to generate the recommendations to one user. This last time is composed by the clustering time (0.68 seconds), creation of recommended mosaics (0.10 seconds), creation of recommended images (0.08 seconds) and creation of recommended overlaps (0.07 seconds). In order to evaluate concurrent queries, the client bot sent to the geoportal microservice 10 and 100 simple of concurrent queries/requests. In the first experiment was observed that, on average, each request was made in 0.75 seconds, whereas when the number of clients was 100, the service time average per query was 0.94 seconds.

When end-users create mosaics and overlapped EOPs on-the-fly, the response time is the metric to be observed because in this type of EOPs the geoportal returns several results to end-users. This time depends on the time spent by a geoportal slave to serve the queries sent by the end-users through their browsers, the volume of results found by the geoportals and the type of shape used to create EOPS based on spatial/temporal parameters. In an experiment, 24 real users sent different shapes to build mosaics and overlapped EOPs and both cases we measured the mean response time observed by these users for each experiment.

Table 1 shows the number of results obtained when the coverage radius of search is changed by the end-users for each catalog considered in this evaluation (Aqua, Terra and LandSat5). For each catalog, we asked end-used for changing the radius until all the raster (individual) images were obtained as a result.

Figure 6 shows, in the vertical axis, the response time in seconds spent by the geoportal to show results of EOPs such as raster images, overlaps, and mosaics when end-users select one, two and three catalogs of satellites (horizontal axis) in the geoportal web page

Table 1 Number of results for different coverage radius of search

# Individual (raster)	# mosaics	# overlaps	Coverage radius (km)
LandSat			
111	29	5	100.421
324	33	13	329.953
1261	35	103	1000.595
1662	36	254	3987.391
Terra			
203	52	42	500.466
656	58	158	1003.251
1805	61	343	2857.952
Aqua			
41	17	22	500.749
406	31	77	1000.302
1033	32	181	2417.834

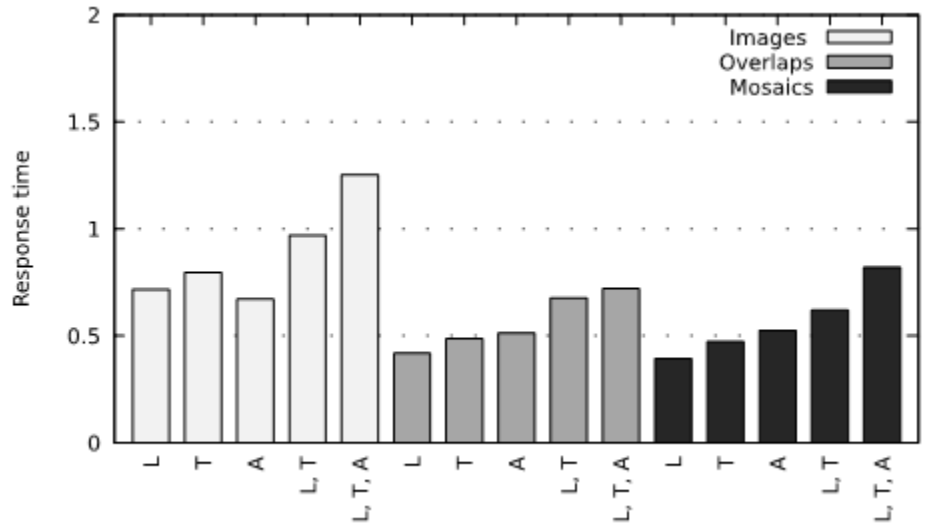

Figure 6: Response time of geoportal serving queries to different satellite sensors: LandSat5(L), Aqua(A) and Terra(T)

Basically, Figure 6 shows three effects affecting the service experience of end-users. The first one is the size of the data source used by the searcher of the geoportal. This means the number of results increases when increasing the number of EOPs in a given catalog, which is increased end-users choosing more sensors in each search. The second effect detected in these tests was that, as expected, the shape chosen by end-users affects the number of results offered by the geoportal to the end-users. For instance, it was observed shapes based on polygons produce increased response times in comparison with other shapes (i.e. circle and rectangle) as the geoportal produces more results of EOPs when these end-users select this type of shape.

Figure 7 shows the resultant classification of the 24 participating users in the previously described experiment, which were organized by the geoportal in clusters of preferences (see users with the same color). The circle shape for each user shows the EOPs of the regions recommended by the geoportal to the end-users. As it can be seen, agencies can identify the volume (in terms of EOPs per cluster) and density (the frequency of these EOPs are requested by end-users) of the consumption of the geoportal service end-users as well as their interests and preferences. This enables agencies to automatically create EOPs. It also enables agencies to either distribute EOPs among their users or build EOPs in advance and make them available in the geoportal depending on the end-user profile described by the clustering of the classification.

Figure 7: Classification of users in clusters of preferences

5. RELATED WORK

Geoportals has been deployed for end-users to search for EOPs with specific characteristics by using simple queries and parameters as the acquisition date, coverage coordinates, and other keywords included into metadata associated to EOPs. Different geoportals have been proposed to manage the information about disaster risks [9], as well as for monitoring rice fields by using the satellite imagery and its derivative products [10].

In order that the end-users can find and download EOPs, retrieval mechanisms as web crawlers were proposed to extract metadata from files found on the web [11]. Also, the information can be retrieved directly from geospatial databases, the metadata is parsed extracting the image information and stored in a database [12]. Geoportals also exhibit derivative products [4, 13] from satellite imagery, which add value to raster images and provide more information to the end-user [14].

Solutions are focused on improving the quality of the results obtained by using semantic web tools such as RDFs, ontology, and SPARQL [14, 15]. Nevertheless, there is an opportunity area to improve the end-user experience by analysing the end-users' activity and grouping them based on their preferences [16]. Moreover, improving the deployment and interoperability of geoportals enable agencies to build clusters to improve the response and service times produced by geoportals when showing EOPs (derivate EOPs as well), which are aspects that were included in the geoportal proposed in this paper.

6. CONCLUSIONS

This paper presented the design and development of an interoperable geoportal service for discovery and management of earth observation products. In this service, the geoportal components are encapsulated into virtual containers launched into the cloud by using a microservice scheme to solve issues such as interoperability (with other systems) and implementation (over different platforms).

A search microservice that analyses the preferences of end-users (settings of spatiotemporal and polygon shapes) and builds clusters of users sharing preferences was included into the geoportal for recommending/delivering, in advance, products matching with end-user preferences. The geoportal service also enables end-users to create, on-the-fly by using spatiotemporal parameters.

A prototype of this service was implemented in a private cloud and connected to a satellite imagery repository of an antenna (ERIS) managed by Mexican Space Agency in a proof of concept. Learned lessons and performance assessments are described through an experimental evaluation with real users participation.

ACKNOWLEDGEMENTS

This work was partially supported by the sectoral fund of research, technological development and innovation in space activities of the Mexican National Council of Science and Technology (CONACYT) and the Mexican Space Agency (AEM), project No.262891.

REFERENCES

[1] Bernard, L., Kanellopoulos, I., Annoni, A. and Smits, P. (2005). The European geoportal—one step towards the establishment of a European Spatial Data Infrastructure. Computers, Environment and Urban Systems, 29(1), pp.15-31.

[2] Geoplatform.gov. (2018). Home. [online] Available at: https://www.geoplatform.gov/ [Accessed 21 Apr. 2018].

[3] Gonzalez-Compean, J., Sosa-Sosa, V., Diaz-Perez, A., Carretero, J. and Marcelin-Jimenez, R. (2017). FedIDS: a federated cloud storage architecture and satellite image delivery service for building dependable geospatial platforms. International Journal of Digital Earth, pp.1-22.

[4] Zhang, Y., Wang, B., Zhang, Z., Duan, Y., Zhang, Y., Sun, M. and Ji, S. (2014). Fully automatic generation of geoinformation products with chinese zy-3 satellite imagery. The Photogrammetric Record, 29(148), pp.383-401.

[5] Bernstein, D. (2014). Containers and Cloud: From LXC to Docker to Kubernetes. IEEE Cloud Computing, 1(3), pp.81-84.

[6] Hu, Y., Janowicz, K., Prasad, S., & Gao, S. (2015). Enabling semantic search and knowledge discovery for ArcGIS Online: A linked-data-driven approach. In AGILE 2015 (pp. 107-124). Springer, Cham.

[7] Vazquez-Martinez, G., Gonzalez-Compean, J., Sosa-Sosa, V., Morales-Sandoval, M. and Perez, J. (2018). CloudChain: A novel distribution model for digital products based on supply chain principles. International Journal of Information Management, 39, pp.90-103.

[8] Gonzalez-Compean, J., Sosa-Sosa, V., Diaz-Perez, A., Carretero, J. and Yanez-Sierra, J. (2018). Sacbe: A building block approach for constructing efficient and flexible end-to-end cloud storage. Journal of Systems and Software, 135, pp.143-156.

[9] Molina, M. and Bayarri, S. (2011). A multinational SDI-based system to facilitate disaster risk management in the Andean Community. Computers & Geosciences, 37(9), pp.1501-1510.

[10] Granell, C., Miralles, I., Rodríguez-Pupo, L., González-Pérez, A., Casteleyn, S., Busetto, L., Pepe, M., Boschetti, M. and Huerta, J. (2017). Conceptual Architecture and Service-Oriented Implementation of a Regional Geoportal for Rice Monitoring. ISPRS International Journal of Geo-Information, 6(7), p.191.

[11] Skluzacek, T., Chard, K. and Foster, I. (2016). Klimatic: A Virtual Data Lake for Harvesting and Distribution of Geospatial Data. 2016 1st Joint International Workshop on Parallel Data Storage and data Intensive Scalable Computing Systems (PDSW-DISCS).

[12] Steiniger, S., de la Fuente, H., Fuentes, C., Barton, J. and Muñoz, J. (2017). BUILDING A GEOGRAPHIC DATA REPOSITORY FOR URBAN RESEARCH WITH FREE SOFTWARE – LEARNING FROM Observatorio.CEDEUS.cl. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-4/W2, pp.147-153.

[13] Institute, W. (2018). Interactive Map | Global Forest Watch. [online] Globalforestwatch.org. Available at: https://www.globalforestwatch.org/map [Accessed 21 Apr. 2018].

[14] de Andrade, F., de Souza Baptista, C. and Davis, C. (2014). Improving geographic information retrieval in spatial data infrastructures. GeoInformatica, 18(4), pp.793-818.

[15] Aditya, T. and Kraak, M. (2007). Aim4GDI: Facilitating the Synthesis of GDI Resources through Mapping and Superimpositions of Metadata Summaries. GeoInformatica, 11(4), pp.459-478.

[16] Dareshiri, S., Farnaghi, M. and Sahelgozin, M. (2017). A recommender geoportal for geospatial resource discovery and recommendation. Journal of Spatial Science, pp.1-23.

AUTHORS

Dante D. Sánchez-Gallegos is a master student at Information Technology Laboratory, Center of Research and Advanced Studies of the National Polytechnic Institute, (CINVESTAV), Ciudad Victoria, Mexico. His research areas of interest are data analysis and cloud computing.

J. L. Gonzalez-Compean received his Ph.D. in Computer architecture from UPC Universitat Politècnica de Catalunya, Barcelona (2009). He was a visiting professor at Universidad Carlos III de Madrid, Spain and researcher at Cinvestav, Tamaulipas, Mexico. His research lines are Cloud-Based Storage systems, linguistic archival systems, federated storage networks. His expertise areas are the design of fault-tolerant, adaptability and availability strategies as well as task scheduling and storage virtualization.

Victor J. Sosa-Sosa is a full-time research-professor at CINVESTAV-Tamaulipas - IPN. He has a Ph.D. in Computer Science from Technical University of Catalonia (UPC-Barcelona). He was professor and researcher in the Distributed Systems Group at CENIDET-Mexico until 2006. Participant and responsible of research and development projects funded by the National Council for Science and Technology of Mexico (CONACYT) and private companies. His research interest and specialization areas are related to Distributed Information Systems and Databases. Recently, his work is more focused on distributed information at Web scale, especially on topics related to efficient information search, storage and analysis, integration of Web query interfaces (DeepWeb) and knowledge harvesting.

José Tuxpan-Vargas was born in Puebla, Mexico, in 1982. He received the B.S. degrees in electronic engineering from the Technological Institute of Puebla, in 2005 and the M.S and Ph.D. degrees in telecommunications from Centre for Research and Advanced Studies of the IPN, Jalisco in 2010 and 2014 respectively. His interests include the design and implementation of new algorithms for imagery systems, adaptive signal /image processing, monitoring and interpretation of geophysical data using remote sensing techniques.

Heidy M. Marin-Castro received the Ph.D. degree in Computer Science from Information Technology Lab in Center for Research and Advanced Studies of the National Polytechnic Institute, Mexico (Cinvestav) in 2014, the M.Sc. degree in Computer Science from the National Institute for Astrophysics, Optics and Electronics (INAOE) in 2008 and the B.Sc. degree in Computer Science from the Autonomous University of Puebla. Currently she is a professor researcher at Engineering and Sciences Faculty in the Autonomous University of Tamaulipas, Mexico. Her research areas include Web Data Management, Databases, Data Mining and Information Retrieval.

INTENTIONAL BLANK

SERVERLESS CLOUD COMPUTING: A COMPARISON BETWEEN "FUNCTION AS A SERVICE" PLATFORMS

Víctor Juan Expósito Jiménez and Herwig Zeiner

JOANNEUM RESEARCH
Forschungsgesellschaft mbH, Steyrergasse, Graz, Austria

ABSTRACT

Cloud computing is moving fast and continually progressing. Beyond the microservices architecture, a new paradigm appears to be evolving and complementing it. By using a serverless computing architecture, faster and more reliable developments are possible in several fields such as the Internet of Things, industrial or mobility applications. In this paper, a serverless computing architecture is described and, in addition, a comparison of the most important "Function as a Service" platforms is given.

KEYWORDS

Serverless Architecture, Function as a Service (FaaS), Cloud Computing, Microservices

1. INTRODUCTION

The Internet of Things (IoT) has changed the way applications are designed. Billions of devices will be connected to the network in a near future. The next generation of connected applications has to change the way they have been designed to support this exponential growth of connected devices and their corresponding services.

At the moment, a microservices architecture can fulfill these requirements where the combination of different services is a main issue. In this kind of cloud computing architecture, all tasks and the logic are split in small services, where each is able to serve just one specialized purpose instead of one monolithic component for all purposes. This allows a more versatile development in which maintenance and updates can be done only to the needed components without changing the whole core system. Every service is isolated and all information and all controllers are accessible through external APIs. The popularization and the standardization of RESTful APIs give developers the ability to use some external services which can take too many resources for a self-implementation. Moreover, reliability, security or scalability are some of the few constrains modern applications have to handle.

Going a step forward, what if there was a way to build a pure native cloud application in which clients (i.e. mobile applications) make all requests and handle the logic of the result without any central server? This type of design is called a serverless computing architecture. This concept will

Dhinaharan Nagamalai et al. (Eds) : ITCSE, WiMo, ICAIT, ICDIPV, CRYPIS - 2018
pp. 15–22, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80702

be possible when more cloud services are available. In addition, cloud computing services have grown exponentially in the last decade. For example, Amazon Web Services (AWS) and Microsoft Azure are constantly updating their cloud services to offer more possibilities to satisfy all requirements that developers could be needing to design and develop applications. In this paper, a detailed analysis of the current status of serverless cloud computing architecture and Function as a Services (FaaS) platforms is given.

The next section gives an overview of existing related work in the field of cloud computing architectures. Section 3 explains the basics of serverless architecture and Function as a Services (FaaS) platforms. In section 4, a brief description of each provider is given. The results of the evaluation are discussed in detail in section 5. Finally, section 6 concludes this work.

2. RELATED WORK

There are many publications which deal with modern architectures and their usage to build cloud-based applications in different fields. Adrian Cockfort [1] explains how modern cloud platforms help companies to speed up the developments and why Netflix decided to move to cloud services to increase the availability and the reliability of the platform. Alan Sill [2] gives some guidelines for designing an application by using a microservices architecture. The authors describe in [3] how the microservices architecture can be used to build a web of things platform. This platform uses some of the principles that define a serverless architecture. Moreover, the authors in this paper are focused on the time constraints and the effect of the delays between microservices on the platform behaviour [4] describes in detail the process and the advantages of using this kind of architectures for an Internet of Things application. Finally, [5] characterizes the implementation of microservices for critical applications in which the sensitive data is handled, integrity and confidentiality are key points.

An important point to consider when designing a cloud application is the cost of the different external services. Andy Singleton [6] discusses how microservices can help companies to avoid errors when a program is too big to be successfully maintained, and the costs that can be saved once a microservices architecture is implemented. He also explains in which cases moving to a microservices architecture makes sense. [7] explains all parameters and costs that developers have to keep in mind when building resilient cloud services.

3. SERVERLESS COMPUTING

A serverless architecture expands the concept where developers do not have to worry about a central server. Some scenarios where this kind of architecture can be successfully implemented are shown in book [8]. Developers can design a complete application with the combination and the communication between third-party services, native cloud services and self-enabled components that give them the flexibility to choose the best service which fulfills the requirements of the application as well as faster development, since an existing service can be instead of developing a one.

On the other hand, the usage of third-party services makes development more dependable on external conditions which are not under the developer's influence. For example, if a provider changes the API of a service, the developer has to modify this aspect on the application or the latter will not be able to connect in future releases. Moreover, all major companies such as

Microsoft or Amazon, which already have a complete cloud services suite, are trying to integrate all services to their own platforms. Therefore, the migration from one provider to another can also be difficult for the interconnection of services from different providers.

In the context of a serverless architecture, two scenarios can be identified:

Backend as a Service (BaaS): Usually a third-party component which provides complete service to the application. In this case, all of the business logic is carried out by the service and the client just gets the results of the task.

Function as a Service (FaaS): The service just runs a small piece of code and gives the result of the function to the client.

In this paper, the main focus is on the analysis and the comparison of FaaS platforms. The usage of FaaS has some advantages in comparison to other types of applications. In traditional applications, developers have to consider all aspects of the environment in which the code is running, like hardware or operating system versions. Once a FaaS platform is used, this procedure is completely transparent to them. Furthermore, no additional time is needed to consider the amount of resources the code needs to be successfully executed on a computer. For instance, the application's design keeps in mind average requests per second to assure the needed hardware is able to handle them with a minimum performance. At this point, there are always some request peaks that may compromise the performance of the whole system due to a lack of resources to handle such an unexpected amount of requests. When it comes to this kind of situation, a FaaS platform comes in handy since the platform automatically handles these request peaks in a transparent way to guarantee a good performance. So developers only have to worry about the application performing well.

When comparing FaaS and PaaS (Platform as a Service), both seem similar at first glance, but on closer look, one separating key aspect becomes evident. According the article [9], Mike Roberts explains that the main difference between both services is the scalability. When PaaS is used, the developer has to keep in mind the resources in case the application receives a request peaks. This process is transparent, however, if FaaS is used.

4. PLATFORMS

This section gives an overview of the principal FaaS platforms as well as a description of the functionality and the special features each one implements.

4.1. AWS Lambda

The Amazon cloud provider was the first to offer a FaaS platform at the end of 2014. Lambda is offered as an isolate service from other AWS services and the price is calculated on the basis of two parameters: the number of executions and the execution computing time according to a defined memory. Amazon also includes a free tier per month before charging costs. AWS Lambda is able to run code of the following programming languages: Java, JavaScript, Python and, since November 2016, C#. The service is strongly integrated with the Amazon Web Services and functions can be triggered by other services too, like Kinesis, S3 or DynamoDB. Unfortunately, a HTTP trigger is limited to the Amazon API Gateway Service which may add

complexity and some delays. One of the flaws of Lambda is the way in which dependencies are handled. There are no configuration files in which developers can version a function or define its dependencies. Therefore, a complete package must be uploaded every time the function changes.

Table 1. Comparison between Function as a Service platforms.

	AWS Lambda	**Azure Functions**	**Google Cloud Functions**	**IBM OpenWhisk**
Resease Date	Nov 2014	Nov 2016	Beta	Feb 2016
Price	$0.00001667/GB-s and $0.20 million execs	Azure subscription plan or $0.000016/GB-s and $0.20 Million Execs	$0.0000025 GB-s and $0.40 Million Execs	IBM Bluemix plan or $0.000017/GB-s
Monthly Free Tier	400,000 GB-s and 1 million execs	400,000 GB-s and 1 million execs	400,000 GB-s and 2 million execs	400,000 GB-s
Maximum time to execute	300 seconds	N/A	540 seconds	300 seconds
Compatibility	Java, JavaScript, Python	JavaScript, Python, PHP, C#, F#, bash, batch	JavaScript	JavaScript, Python, Swift, Docker
Available memory usage	128MB-1536MB	128MB-1536MB	128MB-2048MB	128MB-512MB
HTTP trigger	API Gateway	Native	Native	Native
Open source	No	Yes (Runtime)	No	Yes (Runtime)

4.2. Microsoft Functions

Previously known as a part of Azure WebJobs, Azure Functions was released in November 2016 as an isolated service within the Azure cloud suite. With the release, the related runtime was also published as open source under MIT license and is available in its GitHub repository [10]. The fees of the service follows the same calculation rates as AWS Lambda and it also includes a free grant with the same features. Unlike Lambda, Azure Functions service can be purchased as a pay per demand model or as a part of an Azure subscription plan. Besides the compatibility with common programming languages such as C#, F#, and JavaScript; Functions is also able to execute scripts that use the Windows Command Line, the Power Shell syntax as well as the more common PHP and Python functions which provide more flexibility to the developers; unfortunately, these are still in an experimental status. Moreover, the functions are accessible through HTTP without using any API gateway.

4.3. Google Cloud Functions

Like the other big cloud players, Google has also developed its own FaaS platform called Google Cloud Functions. This service is still a beta release, so its possible functionality and features will be extended in the release version. Unlike other providers, Google Cloud Functions currently supports only JavaScript. On the other hand it is different to the other services. The service allows the usage of more memory, in this case, a maximum of 2GB, and also the free tier gives 1 million executions more than the Azure Functions or AWS Lambda. Moreover, the maximum time of execution for a function increases until 9 minutes. Another interesting aspect of Google Cloud Functions is the possibility of calculating the bill of the service by using GHzseconds rather than the otherwise commonly used GB-seconds.

4.4. IBM OpenWhisk

Starting in February 2016, IBM OpenWhisk was the first open source FaaS, followed by Microsoft with its Azure Functions later. The runtime code is available at GitHub[11] under Apache 2.0 license. IBM OpenWhisk has two price modes, per demand or associated to an IBM BlueMix plan. OpenWhisk supports JavaScript, Python and also adds a compatibility with Swift, the Apple programming language. So far, it is the only service that implements this language and it could be interesting for iOS developers. Another feature that excels this service is the possibility of using Docker containers to run any function or the native implementation of artificial intelligence service, IBM Watson.

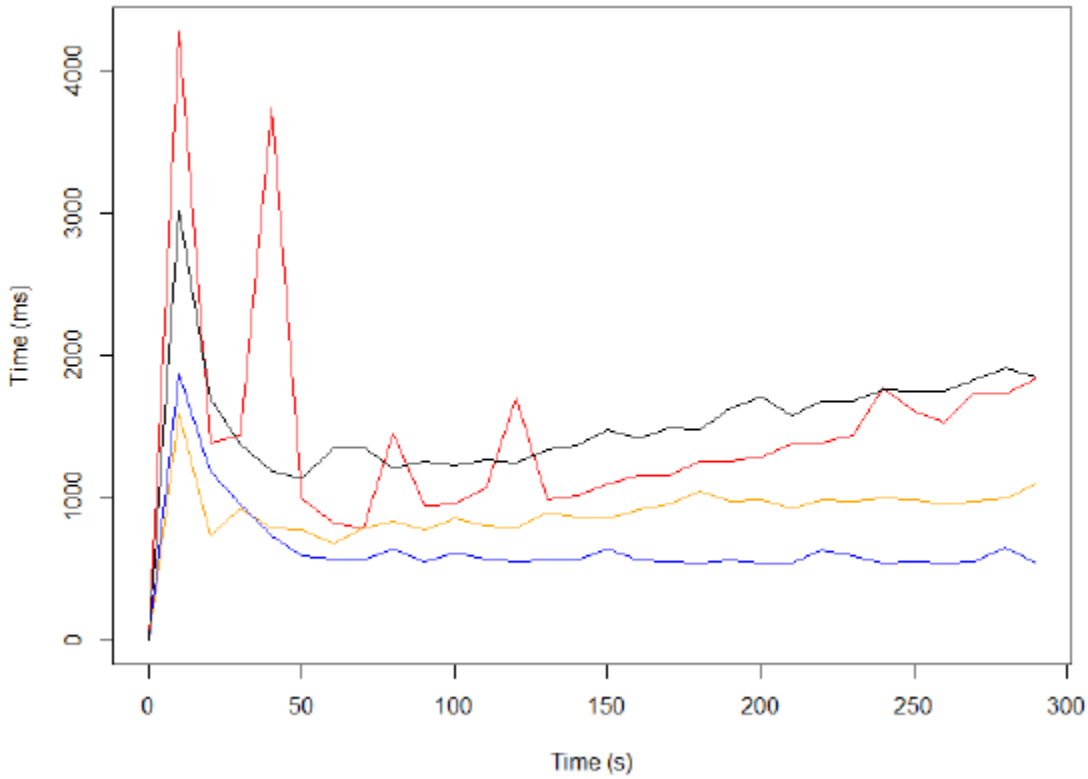

Figure 1. Mean RTTs when the response includes the random generated string by the function. Azure Functions: blue line, IBM OpenWhisk: black line, AWS Lambda: orange line, Google Cloud Functions: red line

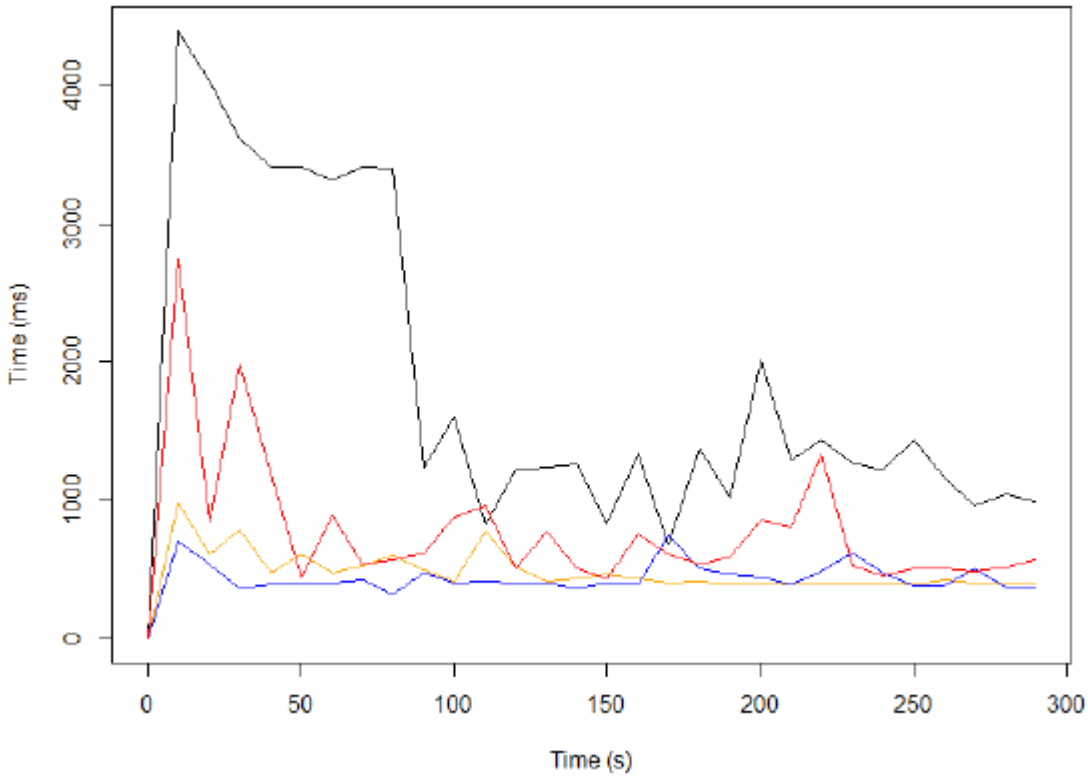

Figure 2. Mean RTTs when the response has an empty body. Azure Functions: blue line, IBM OpenWhisk: black line, AWS Lambda: orange line, Google Cloud Functions: red line

5. EVALUATION

Table 1 summarizes all features of all services that have been described in this paper; for a more detailed evaluation, a simulation has been done. Therefore, a small JavaScript function which generates some computer processing by the calculation of random strings was programmed. JavaScript, being the most supported programming language, was chosen as function programming language. Besides, one point has to be considered during the simulation. At the beginning, we considered as one of the simulation´s parameters the usage of the closest server to our location (i.e. Austria). Unfortunately, OpenWhisk and Google Cloud Functions are still not available in Europe, so all platform simulations had to be done by using servers located in the United States (US). Azure Functions, Google Cloud Functions and IBM OpenWhisk are located in the south central US, Amazon Lambda in the east.

The simulation consists of a client which sends HTTP requests to trigger the FaaS function. Starting with no requests, the requests rate per second increases every ten seconds by adding ten more requests per second. At the end of the simulation (after five minutes), the request rate is 300 req/s which represents a linear increase rate. The reason for increasing the simulation requests is to see how the platform scales when more simultaneous requests trigger the function. The figures show the mean time of the Round-Trip Times (RTT) the platforms give in two determined cases. Figure 1 depicts the mean RTTs when the response includes the random generated string by the JavaScript function. Figure 2, on the other hand, shows the mean RTTs when an empty response

body is given. These two cases have been chosen to see how the body (approximately 80KB) response affects to delays.

All four platforms depict stable behavior with just some peaks at the moment before the platforms scale themselves. The peaks are especially visible in the case of Google Could Function when it scales itself as being clear. Azure Function gives the best performance, both with and without response body, which is always around 500ms; on the other hand, IBM OpenWhisk provides the worst values which are always higher than 1000ms. Amazon Lambda provides a good RTT performance when there is no response body in the request, but it is worse once the randomly generated string is included in the response body and more requests are sent. This may be the result of the required configuration of the API Gateway to enable the HTTP trigger on the Lambda platform. Google Cloud Functions does not present a good performance once the body is included in the response, but the behavior is similar to Azure or Lambda when the response body is empty. One point is especially interesting: all figures show a peak when the simulation starts but stable themself once the simulation continues, which could mean that the platforms need a small period of time to initiate.

According to all analyzed data, the Azure Function platform offers a better performance and supports more programming languages. Moreover, the price is the same compared to the closest competitor, AWS Lambda. IBM OpenWhisk yielded a worse performance but is cheaper than the other platforms and the compatibility with Docker containers, Swift and IBM Watson could be interesting features for some developers. Finally, Google Cloud Functions presents a middle point in features, performance, and price to the others competitors.

6. CONCLUSION

The paper gives an overview and analyses on the current state-of-the-art of serverless cloud architecture. Afterwards, a comparison between the most used FaaS platforms was described in which some parameters were analyzed. These parameters will be used in the future to choose the most suited platform for our requirements. This paper is a basis which will be continuously updated as an extended live document to reflect the changes that may occur in the future on these platforms. The next step will be a more technical comparison of the platforms in a real life scenario, where more precise requirements in delays, programming languages and price are given.

ACKNOWLEDGMENT

This research was partly funded by the Comet-Project DeSSnet- Dependable, secure and timeaware sensor networks. The K-Project DeSSnet is funded within the context of COMET – Competence Centers for Excellent Technologies by the Austrian Ministry for Transport, Innovation and Technology (BMVIT), the Federal Ministry for Digital and Economic Affairs (BMDW), and the federal states of Styria and Carinthia. The programme is conducted by the Austrian Research Promotion Agency (FFG).

REFERENCES

[1] S. Tilkov, “The modern cloud-based platform,” IEEE Software, vol. 32, no. 2, pp. 116–116, Mar 2015.

[2] A. Sill, “The design and architecture of microservices,” IEEE Cloud Computing, vol. 3, no. 5, pp.76–80, Sept 2016.

[3] H. Zeiner, M. Goller, V. J. Expósito Jiménez, F. Salmhofer, and W. Haas, “Secos: Web of things platform based on a microservices architecture and support of time awareness,” e & i Elektrotechnik und Informationstechnik, vol. 133, no. 3, pp. 158–162, 2016. [Online]. Available: http://dx.doi.org/10.1007/s00502-016-0404-z

[4] B. Butzin, F. Golatowski, and D. Timmermann, “Microservices approach for the internet of things,” in 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), Sept 2016, pp. 1–6.

[5] C. Fetzer, “Building critical applications using microservices,” IEEE Security Privacy, vol. 14, no.6, pp. 86–89, Nov 2016.

[6] A. Singleton, “The economics of microservices,” IEEE Cloud Computing, vol. 3, no. 5, pp. 16–20, Sept 2016.

[7] B. Wagner and A. Sood, “Economics of resilient cloud services,” in 2016 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C), Aug 2016, pp. 368–374.

[8] P. Sbarski and S. Kroonenburg, ”Serverless Architectures on AWS”. Shelter Island, New York: Manning Publications Co.,2017.

[9] M. Roberts. (2016, Aug.) “Serverless architectures”.ThoughtWorks, Inc. [Online]. Available: http://martinfowler.com/articles/serverless.html [Accessed: 02- 05- 2018]

[10] Microsoft Corporation, “WebJobs.Script”. [Online]. Available: https://github.com/Azure/azurewebjobs-sdk-script. [Accessed: 02- 05- 2018]

[11] IBM Inc., “OpenWhisk”. [Online]. Available: https://github.com/openwhisk/openwhisk. [Accessed: 02- 05- 2018]

AUTHOR

Víctor J. Expósito Jiménez studied telecommunications engineering at the University of Seville (Spain) and Graz University of Technology (Austria). Since 2012, he works at JOANNEUM RESEARCH in which his main topics have been related to Internet of Things and Cloud Computing. During these years he was one of the people in charge of the development of a 'Web of Things' application platform, designing the architecture and core components. Moreover, he has worked in several projects in different topics such as RFID, wireless sensor networks, and robotics.

Herwig Zeiner is key researcher for Industrial Internet. He was project manager and coordinator for JOANNEUM RESEARCH of several projects. His research interests are time-aware data-analytics in industrial applications, distributed service-oriented and event driven applications, data driven applications by using stream analytics, as well as intelligent feature extraction techniques and different data mining-methods for event streams. He is also an evaluator of the European Commission.

A REVIEW OF TECHNIQUES FOR POSITIONING IN WLAN WITH LIMITED DATA

Vitaly Romanov and Giancarlo Succi

Innopolis University, Innopolis, Republic of Tatarstan, Russia

ABSTRACT

Traditionally, positioning in WLAN was associated with some issues. The presence of the multipath forced researchers to resort to fingerprinting based positioning techniques that inherently require extensive site surveying and the abundance of reference signals. However, if we consider WiFi data for the task of crowd analysis, the data should be collected on the side of the network provider and, in this case, it is usually scarce. Thus, methods that require fewer reference signals for positioning are needed. This paper provides the comparison of WLAN based positioning methods that can operate with a single AP.

KEYWORDS

Localization, Wireless, WiFi, Mobility Tracking, Mobile Users.

1. INTRODUCTION

One of the applications of positioning systems is crowd mobility analysis that can be used to detect common patterns in crowd movement, irregular commotions in the area, etc. This task largely relies on the ability to estimate the locations of people in the crowd. Several technologies are currently available to provide reliable location information. The most widely known - Global Positioning System (GPS) - is successful in doing this outdoors. However, it is reported that positioning precision suffers a lot in city landscapes, and requires additional reference signals or calibration [1]. A local area positioning system can provide one of the solutions for this problem.

One of the requirements of a successful and efficient positioning system is the ease of creating the infrastructure. The popularity of GPS is caused by its availability worldwide, and this trait is hard to achieve for other positioning systems by design. Ideally, a local positioning system would be very cheap to set up and sustain. For this reason, there are attempts to adopt cellular or wireless local area networks (WLAN) as the basis for a positioning system that would enable localization indoors or in dense urban environments [13]. Cellular data is ideal for the task of crowd mobility tracking (in term of coverage), but it provides very rough granularity. WLAN data does not suffer from this issue.

Dhinaharan Nagamalai et al. (Eds) : ITCSE, WiMo, ICAIT, ICDIPV, CRYPIS - 2018
pp. 23–33, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80703

The major problem for crowd analysis task is that the data should be collected on the side of the network provider and, in this case, it is usually scarce. The most popular method for positioning with fingerprints requires several reference signal to achieve reasonable precision. Such abundance of information is often unavailable on the network service provider side. Thus, methods that require fewer reference signals for positioning are needed.

Many of classical positioning approaches that use cellular networks or WLAN are based on triangulation and trilateration techniques, which are not applicable when less than three measurements from different base stations are provided. Another class of methods, based on the distribution of received signal strength (RSS), made its appearance in the last decades due to the availability of vast computational resources. Although more computationally intensive, these new methods allow achieving reasonable positioning accuracy in environments subject to multipath propagation.

RSS based techniques proliferated in the area of WLAN-based positioning where the measure of received signal strength is provided by Received Signal Strength Indicator (RSSI). WiFi networks are most common for indoor spaces, which simplifies the application of RSSI positioning techniques that are usually referred to as fingerprinting methods. Indeed, the indoor area is constrained by the borders of the premises, and the rest of the surrounding environment lies outside the consideration. When applied to a larger scale positioning, namely for cellular networks, the computational diffculty of signal strength based methods explodes due to an inherently large size of fingerprint database, and the need for site surveying.

Indoor WLAN-based positioning with RSSI is usually performed with measurements from multiple access points (APs). However, there are situations when the area is covered only by a single AP, or information from only a single AP can be utilized appropriately (i.e., the case of network-side data collection). In these scenarios, techniques like triangulation and trilateration are hardly applicable. The contribution of this paper is in providing the comparison of WLAN based positioning methods that requires can operate with a single active AP.

The rest of the paper is organized as follows. In section 2 we overview related work. In section 3 we formulate the problem of positioning in WiFi networks and discuss signals used for positioning and limitations that emerge when positioning with a single AP. In section 5 we classify existing techniques for positioning with one AP. Section 6 concludes the paper.

2. RELATED WORK

Positioning in WLAN is related to other forms of wireless positioning, more precisely, positioning with radio signals. Liu et al. (2007) provided a general review of methods for localization in wireless networks and performed their comparison concerning accuracy, precision, complexity, scalability, robustness, and cost [8].

Farid et al. (2013) considered only indoor positioning techniques [3]. They included recent work and performed the comparison of positioning methods using other criteria, such as coverage provided by the system, power consumption, and susceptibility to multipath.

Quoc Duy Vo and Pradipta De (2016) compared existing wireless positioning techniques for outdoors concerning accuracy and power consumption [13].

Until recently, there were two types of information that a client or an AP could measure during communicating in WLAN. The first is the RSSI and the second is temporal delay. With the introduction of MIMO (Multiple Input Multiple Output) in IEEE 802.11 standard, it became much simpler to obtain the information about the direction of the incoming signal.

The initial attempt to create indoor positioning system with WiFi could be classified as time-based or signal strength-based. Although time-based methods often do not allow to use off-the-shelf hardware, Makki et al. (2015) explored the opportunity of positioning using time measurements in WiFi and discussed existing ways to overcome the issues with resolution and the prospective opportunities of time-based positioning in WLAN [9].

Suining He and Gary Chan (2016) considered only fingerprinting methods for localization [4]. They noted the use of spatiotemporal patterns, collaborative localization, and motion assistance in recent papers. Additionally, they reviewed different techniques for reducing the site survey, calibrating the system for heterogeneous devices, and performed the comparison of many methods using various criteria.

Although RSSI-based methods deal with the problem of multipath, they require the knowledge of space configuration and the presence of signal strength map, which takes a lot of time and resources to collect. Mahtab Hossain and Wee-Seng Soh (2015) decided to look into calibration-free indoor positioning methods, which do not require extensive map collection or even the precise knowledge of the floor plan [5]. For comparing these calibration free methods, they considered map requirements, the necessity of the initial location fix, the amount of user participation, and the need for additional sensors, such as inertial sensors.

The difference of our review is that we consider only positioning techniques that require only a single active AP to localize the client. Intuitively, if a method achieves reasonable localization error with a single AP, it's results are likely to improve when scaled to several APs. Papers covered in this review were not considered in previous surveys. Our contribution is in aggregating the information about this subtype of positioning methods and providing their comparison.

3. PROBLEM FORMULATION

3.1 Positioning Procedure

The goal of positioning is to identify the location of a user relative to a set of base stations (BS). In the case of WLAN-based positioning, the user is a network client, and APs serve as BS. Conventionally, two types of positioning approaches are distinguished: deterministic and probabilistic [4]. Deterministic algorithms use some similarity or distance metric to compare observed signals with a prerecorded database. In this case, the positioning procedure becomes an optimization problem

$$\hat{l} = \mathrm{argmin}_l SIM(\mathbf{s}, \mathbf{s}_l)^b$$

where *SIM* is a similarity function, and b = 1 when *SIM* (s,s') increases as s' → s and b = -1 if otherwise. Here, *l* denotes the location, and s_l - the signal associated with the location.

On the other hand, the probabilistic approach employs maximum likelihood (ML) technique where a probability distribution ties together the signal space and locations on the map and the goal is to maximize the conditional probability

$$\hat{l} = \text{argmax}_l \mathbf{P}(l|\mathbf{s})$$

These two approaches constitute the basis for building positioning algorithms and the papers considered further use either of them.

Besides choosing the location estimation method, it is essential to select an appropriate set of signals. In the next section, we will discuss the types of measurements that are used for positioning in WLAN.

3.2 Signals for positioning in WLAN

Measures that are used for positioning in WLAN include RSSI, time or phase, the angle of arrival (AoA), and channel state information (CSI). In this paper, we refer to those as location signals, and their properties are described below.

RSSI measurements are prone to high variance especially in indoor environments [2]. Although RSSI was initially adopted for positioning due to its indifference to multipath, latest work shows that it suffers from non-stationary variance [6]. Nevertheless, positioning with RSSI in WLAN is the most common technique to date.

Positioning in WLAN can be done using time or phase measurements. The simple idea is that these measurements can be converted into the distance that is further used to localize a user. Although less sensitive to noise, time intervals are harder to measure using off-the-shelf WLAN adapters due to internal clock resolution and the radio bandwidth of the received signal [9]. It was estimated that with the standard bandwidth of 40 MHz the inaccuracy of time measurements results in localization errors around 7 meters [11]. This number can be improved by applying special techniques.

CSI describes the state of the communication channel and provides rough information about channel impulse response. CSI is represented by phase and amplitude of the received signal in the frequency domain. The IFFT of CSI gives the notion of delays of different multipath components. Recent work treats it as a promising localization signal [2]. Currently, CSI is not available by default, and the driver for a wireless adapter should be modified to obtain the access to these measurements.

Methods for positioning described in this review use at least one of the signals above to infer user's location and some of them use their combination to improve the precision.

3.3 Positioning with a single active AP

In general, the level of confidence for estimated location is low when positioning with a limited number of APs. For an indoor environment, this problem can be addressed by considering the joint distribution of RSSI and the location on the map. The presence of obstacles on the way of WiFi signal propagation, like walls, create an irregular RSSI pattern across the area. Some patterns make it easier to infer the true location. The distribution of average RSSI across space is

called coverage map. Techniques that utilize such coverage maps are usually called scene analysis techniques [8]. It is worth noting that the irregularity of a coverage map highly depends on the configuration of the environment, and in open spaces such maps do not provide significant benefits over other less memory intensive techniques such as statistical propagation models.

Another approach is related to the use of triangulation or trilateration. The position of a client can be identified when sets of distances and/or AoAs are known. With the introduction of MIMO in the current IEEE 802.11 standard, each compatible AP is equipped with multiple receiver inputs. This receiver antenna array can be used to infer the relative position of a user and the AP. For simplicity, we further refer to triangulation and trilateration techniques under a common term of triangulation.

We were able to discern three classes of positioning techniques with a single AP: (i) triangulation (ii) triangulation and dead-reckoning (DR) fusion (iii) scene analysis and dead-reckoning fusion.

The first class refers to techniques that rely on the one-time measurements, i.e. the location can be inferred from a single sample of all necessary parameters. The second one includes methods with similar positioning principles but relies on dead-reckoning for obtaining the final location estimate. The third - uses RSSI and coverage map together with dead-reckoning for localization.

4. COMPARISON CRITERIA

We choose to compare methods presented here using the following attributes. First, we distinguish the use of different positioning approaches, which define the set of used localization signals and processing techniques.

Often, the positioning system becomes harder to implement when there is a need for additional sensors. Therefore, we try to identify the extent to which a particular positioning procedure uses other measurements.

Some approaches must take advantage of low-level information from the physical layer of the communication protocol (PHY). This is often associated with particular modification and requires additional work during deployment.

The sampling rate (SR) is often crucial for positioning, especially when deadreckoning is used. High SR is associated with higher energy consumption and communication overhead.

The amount of user involvement in the positioning process differs for presented positioning methods. The necessity of additional computation on the side of the user requires both energy resources and specialized software, hindering seamless deployment of the positioning system.

One of the most critical factors for deployment is the necessity of extensive site survey. While it is nil for triangulation techniques, the presence of fingerprint database is the requirement for scene analysis methods.

Further, we are going to compare different techniques concerning these parameters. All information is taken from the reports produces by authors of corresponding methods, and we specify when we make our own notes and conclusions.

5. POSITIONING TECHNIQUES

5.1 Triangulation

The class of triangulation techniques includes methods that rely only on explicitly measured parameters and exclude the need for tracking. Thus, these techniques allow for positioning at any moment of time without the knowledge of previous locations. Also, they use only measurements for distance and angles.

CHRONOS The positioning system, named *Chronos*, designed by Vashisht et al. (2016), is based on commodity wireless adapters and allows localization with a decimeter level of precision [12]. It is claimed to be the first system that performs positioning with such accuracy without external sensors. The core idea of the method is to estimate distances between antennas of the AP and a client using absolute ToF.

Chronos uses commodity hardware Intel 5300 with a modified driver. The use of ToF is associated with a series of issues that were discussed in the section 3.2 but authors were able to find a way mitigate these limiting constraints.

Authors substantiate the claim that the phase of the zeroth sub-carrier in OFDM spectrum of WLAN transmission is equal to the true value of the phase, which allows eliminating additional phase shift due to packet detection delay. Thus,

ToF can be computed from the expression [12]:

$$\phi_{i,k} = -2\pi f_{i,k}\tau \mod 2\pi$$

where τ is the ToF, and $f_{i,k}$ is the frequency of the sub-carrier k in the frequency band *i*, and $\phi_{i,k}$ is measured by the receiver. Obviously, such an equation has an infinite number of solutions for τ.The unique solution can be found from a system of equations that uses frequencies for different channels. For this reason, the localization procedure involves channel sweep.

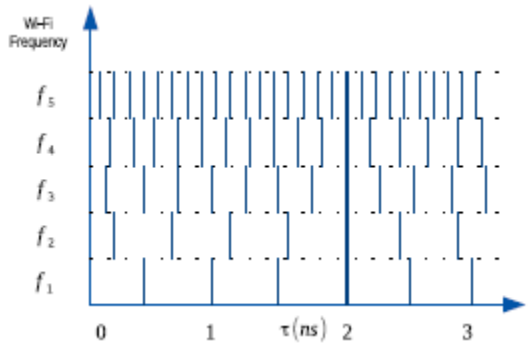

Fig. 1. Example demonstrates the process for solving the ToF based on measured phase. In each channel, there is a set of periodic solutions. The answer corresponds to the time when solutions for all the channels match [12].

The problem of multipath is dealt with using channel state information (CSI), which allows separating the component of the direct path from the rest. The final location of a user is found by solving a quadratic optimization problem. The goal of such optimization is to minimize the difference of distances between a user's assumed location and the AP, and distance estimations calculated from ToF measurements.

In the case of *Chronos*, the AP should have the modified driver installed. This driver provides access to physical layer information and enables the AP to drive frequency hopping procedure. The distance is estimated from the system of signal phase equations and does not require an onboard high-precision clock. In the case of LOS, 95% of ToF estimation error is contained within 1.96 ns, and the positioning error is under 2m in 90% of cases. One of the essential features of *Chronos* is its ability to produce confident location estimation after a single scan.

The success of Chronos arises the question of the impact of this positioning procedure on the network throughput. It was shown that single localization attempt that involves multiple channel sweeps reduces throughput of the network by approximately 5 MBits/s for the period of localization, which lasts about 84 ms.

Triangulation with passive sensors Another triangulation approach was proposed by Kraxberger et al. (2010) [7]. They developed a method that allows identifying the location without active collaboration from a user. In their scheme, there is a single active AP that communicates with the client, and several passive sensors. The job of sensors is to monitor the wireless channel and intercept packets transmitted by the client. The RSSI of intercepted packets carries the information about the distance to the client, which can be extracted using Two-Way-Ground propagation model. After distances between the client and three of the sensors are estimated, triangulation is applied.

This technique does not require any modification of hardware or protocol logic and can be used with any off-the-shelf wireless adapter and AP. However, this method relies only on the statistical propagation model, which can be highly imprecise in certain environments. Thus, the effect of multipath is modeled only approximately. The authors do not evaluate the average accuracy of their method, making it difficult to perform the comparison with different techniques.

5.2 Scene analysis and dead-reckoning fusion

Zaruba et al. (2007) presented a method for indoor localization with a single access point that uses both scene analysis and dead-reckoning [14]. By combining these two approaches, they were able to achieve average positioning accuracy of 2m.

Before the positioning procedure can be used, one needs to calculate signal coverage map. To achieve this goal, the accurate floor plan is obtained, and then parametric ray-tracing is performed. By simulating the propagation of radio waves, one can predict the observed RSSI at a given point on the floor plan. To simplify and enhance the ray-tracing process, the passing and reflection coefficients are treated as unknown parameters. Their values obtained by numerically solving the minimization problem where the average squared difference between observed and estimated RSSI measurements serve as the cost function.

Positioning with fingerprints generated in such a way produces a set of possible locations that is too large in general. However, the use of dead-reckoning together with analytical movement model allows filtering the most unlikely positions of users. In this method, no additional sensors on the device are used, and analytical movement model is employed. The consecutive samples of this model follow Markov assumption. User's positions are processed with particle and Bayesian filters to increase the positioning accuracy.

This positioning technique handles the problem of the multipath well and requires no tampering with the WLAN communication protocol. However, the deployment is mostly hindered by the necessity of detailed floor plan. Moreover, the ray tracing procedure is merely an approximation of the real radio wave propagation process and inherently contributes to the final positioning error.

5.3 Triangulation and dead-reckoning fusion

In this class, the difference from the triangulation techniques is that some parameters are not estimated after single sampling is performed. Instead, necessary information is accumulated through the extent of time before the estimation can be produced.

CUPID Sen et al. (2013) designed a method that relies on measurements from client's device and a single AP and does not require extensive site surveying [11]. Their method uses distance estimated from a propagation model, and AoA obtained through dead-reckoning. The median localization error of the proposed technique is 5m. They successfully utilize information about multipath components, provided by CSI, and knowledge of user mobility to implement a system that works on commodity wireless cards.

In the presence of multipath, the value of RSSI can deviate from the expected value, and statistical propagation models are usually unable to reasonably describe this deviation, especially when the direct path component is not the strongest in the delay profile. On the other hand, when multipath components are filtered, the propagation model can be of greater use. Thus, Sen et al. use the energy of direct path (EDP) in conjunction with propagation model to estimate the distance between AP and the client. EDP is easily obtained once CSI is available**.**

The AoA is estimated with the facilitation of dead-reckoning. The true AoA cannot always be determined even when an AP with MIMO capabilities is used. In some scenarios, the strongest energy is contributed by a multipath component, drawing conventional methods for determining the direction of arrival of the strongest signal useless. A pseudo-spectrum is usually constructed outlining the dependence of RSSI from the AoA, and the angle that yields the peak energy value is chosen. If local maxima are considered, the procedure will result in the set of possible angles that needs further refinement. The possible solution is to use movement tracking with accelerometer and estimate traveled distance. As a result, a triangle can be constructed with two sides measured with propagation model, and the third - estimated through dead-reckoning, as shown in Fig. 2. Thus, angles from the set of possible solutions can be tried. After identifying the AoA and the distance, user's position is estimated.

This method introduces a calibration-free approach to local area positioning. It is claimed that the positioning procedure was designed with energy efficiency in mind and probing the client's location with a high rate is not required. On the other hand, the accuracy of DR highly depends on

the sampling rate of the device's internal sensors. The method assumes the presence of direct path in the CSI. Authors assert that measurements of EDP lower than 12 dB are unreliable and should be discarded. This approach can be extended to use several AP with a successful decrease of localization error.

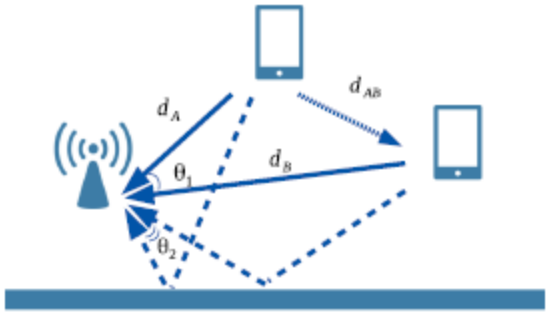

Fig. 2. Angle of direct path corresponds to the triangle with sides d_A, d_B and pAB. When signals from multipath components are tried, the angle does not fit this triangle [11].

SAIL The improvement of the previous approach was introduced by Mariakakis et al. (2014) [10]. The location is determined using distances measured through ToF and DR with the resulting median positioning error of 2.3m. They take advantage of CSI and user mobility to increase the performance of measurements, and device a heuristic for estimating absolute heading.

The ability of the AP to measure the delay between packet departure and ACK signal arrival lies in the foundation of this technique. The precision of time measurements limited by 40 MHz bandwidth is 25 ns. This precision can be enhanced by evaluating the relative time delay of signal arrival to different antennas. Additionally, the incorporation of human mobility and channel coherence property allows for the further decrease of positioning error.

Authors improved DR procedure by increasing the accuracy of the mobile device's inertial system. Driven by the goal to estimate the absolute heading, they develop a procedure to switch between readings from magnetometer and gyroscope depending on the current conditions. As a result, the average heading estimation error dropped from 27.4 to 5.9 degrees.

Eventually, the location was determined by constructing a triangle with two sides estimated using ToF, and the third - using DR. The number of possible triangle orientations was reduced by considering the absolute heading.

In this method, the accuracy of CSI measurements is increased by exploiting the channel coherency. Thus, the measurements are performed several times consecutively. This results in a small communication overhead.

6. CONCLUSION

We have reviewed methods for positioning in WLAN with a single AP. Conventionally, single AP provides only limited information that is not suitable for confident localization. Moreover, techniques for identifying both the distance and the direction to the user are constrained by the presence of multipath.

Table 1. Comparison of techniques for positioning with a single AP. FP - fingerprinting, SR -sampling rate, IMU - inertial measurement unit, PHY - physical layer.

Name	Pos. Method	Extra Sensors	PHY Info.	High SR	User Particip.	Comm. Overhead	Median Error
Chronos[12]	Trilat.	None	Yes	No	Modified driver required	Location fix in 84 ms	64 cm
Zaruba[14]	DR assisted FP	None	No	Yes	Measures RSSI	None	2 m
CUPID[11]	DR assisted trilat.	IMU	Yes	Not required, except for DR	Measures RSSI, performs DR	Low	5 m
SAIL[10]	DR assisted trilat.	IMU	Yes	Not required, except for DR	Measures RSSI, performs DR	0.20%	2.3 m
Kraxberger[7]	Tritat.	Packet sniffers	No	No	None	None	Unknown

The papers described here address the problem of multipath and propose methods that allow precise positioning in areas where it was not feasible before. We compared presented techniques according to the number of criteria that show the trade-off between the complexity of deployment and the complexity of positioning procedure.

Most of the introduced algorithms belong to the class of calibration-free techniques, which presents a significant advantage for deployment. All methods use commodity hardware, and some of them require the access to CSI provided by the driver.

The method presented by Zaruba et al. (2007) and Kraxberger can work with any hardware without limitations. Chronos, CUPID, and SAIL require WLAN standard 802.11n and above in conjunction with a modified driver. Although Chronos has the best positioning accuracy, it requires frequency hopping, that may pose more problems with additional drivers on the user-side and increase channel interference.

Overall, we see that positioning no longer requires the presence of multiple AP. This can enable new opportunities for WLAN based positioning systems and allows for improved accuracy in the case when more than one AP is present.

REFERENCES

[1] Chang, N., Rashidzadeh, R., and Ahmadi, M. (2010). Robust indoor positioning using differential Wi-Fi access points. IEEE Transactions on Consumer Electronics, 56(3):1860-1867.

[2] Chapre, Y., Ignjatovic, A., Seneviratne, A., and Jha, S. (2014). CSI-MIMO: Indoor Wi-Fi fingerprinting system. Proceedings - Conference on Local Computer Networks, LCN, pages 202-209.

[3] Farid, Z., Nordin, R., and Ismail, M. (2013). Recent advances in wireless indoor localization techniques and system. Journal of Computer Networks and Communications, 2013:1-12.

[4] He, S. and Chan, S.-H. G. (2016). Wi-fi fingerprint-based indoor positioning: Recent advances and comparisons. IEEE Communications Surveys & Tutorials, 18(1):466-490.

[5] Hossain, A. M. and Soh, W.-S. (2015). A survey of calibration-free indoor positioning systems. Computer Communications, 66:1-13.

[6] Kaemarungsi, K. and Krishnamurthy, P. (2012). Analysis of wlans received signal strength indication for indoor location fingerprinting. Pervasive and Mobile Computing, 8(2):292-316.

[7] Kraxberger, S., Lackner, G., and Payer, U. (2010). Wlan location determination without active client collaboration. In Proceedings of the 6th International Wireless Communications and Mobile Computing Conference, pages 1188-1192. ACM.

[8] Liu, H., Darabi, H., Banerjee, P., and Liu, J. (2007). Survey of wireless indoor positioning techniques and systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 37(6):1067-1080

[9] Makki, A., Siddig, A., Saad, M., and Bleakley, C. (2015). Survey of wifi positioning using time-based techniques. Computer Networks, 88:218-233.

[10] Mariakakis, A. T., Sen, S., Lee, J., and Kim, K.-H. (2014). Sail: Single access point-based indoor localization. In Proceedings of the 12th annual international conference on Mobile systems, applications, and services, pages 315-328. ACM.

[11] Sen, S., Lee, J., Kim, K.-H., and Congdon, P. (2013). Avoiding multipath to revive inbuilding wifi localization. In Proceeding of the 11th annual international conference on Mobile systems, applications, and services, pages 249-262. ACM.

[12] Vasisht, D., Kumar, S., and Katabi, D. (2016). Decimeter-level localization with a single wifi access point. In 13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16), pages 165-178.

[13] Vo, Q. D. and De, P. (2016). A survey of fingerprint-based outdoor localization. IEEE Communications Surveys & Tutorials, 18(1):491-506.

[14] Zaruba, G. V., Huber, M., Kamangar, F., and Chlamtac, I. (2007). Indoor location tracking using rssi readings from a single wi-fi access point. Wireless networks, 13(2):221-235.

INTENTIONAL BLANK

A METHODOLOGY FOR THE AUTOMATIC CREATION OF MASSIVE CONTINUOUS QUERY DATASETS FROM REAL LIFE CORPORA

Christos Tryfonopoulos

Dept. of Informatics and Telecommunications
University of the Peloponnese, Tripoli, Greece

ABSTRACT

In the information filtering (or publish/ subscribe) paradigm, clients subscribe to a server with continuous queries that express their information needs while information sources publish documents to servers. Whenever a document is published, the continuous queries satisfying this document are found and notifications are sent to appropriate subscribed clients. Although information filtering has been in the research agenda for about half a century, there is a huge paradox when it comes to benchmarking the performance of such systems. There is a striking lack of a benchmarking mechanism (in the form of a large-scale standarised test collection of continuous queries and the relevant document publications) specifically created for evaluating filtering tasks. This work aims at filling this gap by proposing a methodology for automatically creating massive continuous query datasets from available document collections. We intend to publicly release all related material (including the software accompanying the proposed methodology) to the research community after publication.

KEYWORDS

Continuous queries; dataset construction; information filtering; publish/subscribe; information dissemination; profiles;

1. INTRODUCTION

In recent years, content-based *information filtering* (or *publish/subscribe*) applications, such as news or digital library alerts, have gained popularity to help users cope with the information avalanche problem on the Web. In the information filtering paradigm, users -or services that act on users' behalf- *subscribe* to a server with *continuous queries (*or *profiles)* that are expressed in some well-defined language and capture their information needs. When a document is *published* on the server, the continuous queries satisfying the document are found and notifications are sent to appropriate clients. Publishers may be news feeds, digital libraries, or users who post new blog items. Notice that information filtering is very different from information retrieval (as in search engines), in which a user poses a (one-time) query and the search engine executes it only once to retrieve the currently matching documents.

Dhinaharan Nagamalai et al. (Eds) : ITCSE, WiMo, ICAIT, ICDIPV, CRYPIS - 2018
pp. 35–45, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80704

Since a server may handle millions of clients and continuous queries, the filtering problem needs to be solved efficiently by each server. To this end, a number of systems and algorithms that try to solve the filtering problem efficiently for different data models and query languages have been proposed [1, 6, 10, 17, 11, 12, 16, 18, 19, 20]. However, despite all the research in the area, there is an apparent lack of a *benchmarking mechanism* (in the form of a *large-scale standarised test collection* of continuous queries and the relevant document publications) specifically created for evaluating filtering tasks. From our point of view there exist two major problems to be addressed when trying to experimentally evaluate a filtering algorithm: (i) the *document corpus* to be used as publications and (ii) the set of *continuous queries* relative to that corpus. It may not be difficult to collect data to be used as publications since there is a wide collection of document corpora available. It is however, extremely difficult to find continuous queries relevant to a specific corpus except by obtaining *proprietary data* (e.g., from Google Alerts or CNN's news alert system). Notice also that one-time queries, such as those obtained from public releases of major search engines' query logs (like Google BigQuery, Zeitgeist, or the AOL query set) are *inappropriate* for filtering tasks as they typically express a one-time information need, contrary to continuous queries that are used to express *recurrent* and *long-standing* information needs. Finally, other efforts, such as the TREC Filtering Track, are *insufficient* as they contain only a few dozens of *manually* created and curated continuous queries, and cannot live up to the need of modern benchmarking that is in the order of millions (e.g., as in [12, 16, 18, 19, 20]).

Given the above, it becomes clear that the only viable alternative to this lack of standarised benchmarks is to *artificially generate* sets of continuous queries related to the corpus to be used for the evaluation. To the best of our knowledge *this is the first approach in the literature* to provide a *general-purpose methodology* for artificially generating *realistic continuous query datasets* from *actual document corpora* for benchmarking purposes. To this end, our contributions are the following:

- We formally define the query language and data model named AWP supported by our methodology

- We introduce a new corpus of research papers to be used as publications in the filtering tasks. Notice however, that our methodology is general enough to be used with any (attribute tagged) document corpus; here the new corpus is only used as a proof-of-concept for our continuous query creation process.

- We propose a new methodology for creating synthetic user profiles using words and technical terms extracted automatically from the document corpus. To do so we use the corpus at hand to create realistic continuous queries under query language AWP. It should, however, be stressed out that our methodology can be applied to any corpus and any query language similar to AWP

The rest of the paper is organised as follows. Section II gives a brief overview of related work and Section III hints on the model *AWP*, which is used for specifying profiles and documents. Section IV presents the NN corpus, whereas Section V presents our methodology for the creation of continuous queries. Finally, Section VI concludes the paper by providing future research directions.

2. RELATED WORK

Our work relates (at a higher level) to the general area of information filtering efficiency as expressed by a number of systems and algorithms that try to provide scalable information filtering solutions for different data models and query languages. Some of these approaches include the systems XFilter [1], YFilter [6], DFA [10], the Boolean version of SIFT [17], and the agent-based DIAS [11]. Other approaches focus more on the algorithmic aspect by providing efficient tree-based data structures such as [12, 16, 18, 19, 20] for dealing with documents that are free text and profiles that are conjunctions of keywords. To the best of our knowledge the only work that is somewhat relevant to ours is [15], where a corpus of documents (but no continuous queries) is built for adaptive filtering tasks.

Interestingly the evaluation of the XFilter [1] and SIFT [17] is based on a synthetic corpus of documents; XFilter creates them using IBM's XML generator [7] and NITF DTD [5], whereas the creation of continuous queries is also synthetic. Contrary, [10] uses Deterministic Finite Automata to parse a corpus of XML documents and the XPath generator used in YFilter [6] to generate the user profiles. However, the main problem with all these approaches is that they are (i) aimed at a single evaluation and cannot be reused, (ii) based in artificial (and not real document) corpora, and (iii) not freely available to use. Contrary, our approach is general enough to cover many filtering tasks, is based on actual documents to create the continuous query dataset, and will be freely available for use after publication.

3. THE DATA MODEL AWP

In [11] we present the data model AWP for specifying continuous queries and textual resource metadata in information filtering systems. AWP is based on the concept of named attributes with values of type text. The query language of AWP offers Boolean and proximity operators on attribute values as in the work of [4], which is based on the Boolean model of information retrieval.

Syntax. Let Σ be a finite *alphabet*. A *word* is a finite non-empty sequence of letters from Σ. Let V be a (finite or infinite) set of words called the *vocabulary*. A *text value s* of length n over vocabulary V is a total function $s:\{1,2,...,n\}\rightarrow V$.

Let I be a set of *(distance) intervals* $I=\{[l,u]: l,u \in \mathrm{N},\ l \geq 0$ and $l \leq u\} \cup \{[l,\infty): l \in \mathrm{N}$ and $l \geq 0\}$. A *proximity formula* is an expression of the form $w_1 \prec_{i\_1} \cdots \prec_{i\_n-1} w_n$ where $w_1,...,w_n$ are words of V and $i_1,...,i_n$ are intervals of I. Operators $\prec_i$ are called *proximity operators* and are generalizations of the traditional information retrieval operators *kW* and *kN* [4]. Proximity operators are used to capture the concepts of *order* and *distance* between words in a text document. The proximity word pattern $w_1 \prec_{[l,u]} w_2$ stands for "word w_1 is before w_2 and is separated by w_2 by *at least l* and *at most u words*". The interpretation of proximity word patterns with more than one operator $\prec_i$ is similar. A *word pattern* over vocabulary V is a conjunction of words and proximity formulas. An example of a word pattern is $applications \wedge efficient \prec_{[0,0]} data \prec_{[0,3]} fusion$.

Let $\mathring{A}$ be a countably infinite set of attributes called the *attribute universe*. In practice attributes will come from *namespaces* appropriate for the application at hand e.g., from the set of Dublin Core Metadata Elements [21].

A document d is a set of attribute-value pairs (A,s) where A ∈ Å, s is a text value over V, and all attributes are distinct. The following set of pairs is an example document:

{ (AUTHOR, "Christos Tryfonopoulos"),
(TITLE, "Distributed information filtering is ..."),
(ABSTRACT, "In this paper we show that ...") }

A *query* is a conjunction of the form

$$A_1 = s_1 \wedge ... \wedge A_n = s_n \wedge B_1 \sqsupseteq wp_1 \wedge ... \wedge B_m \sqsupseteq wp_m$$

where each Ai,Bi ∈ Å, each si is a text value and each wpi is a word pattern. The following formula is an example query:

AUTHOR = "Christos Tryfonopoulos" ∧
$TITLE \sqsupseteq (distributed \prec_{[0,3]} filtering) \wedge information$

Semantics. The semantics of *AWP* have been defined in [11] and will not be presented here in detail. It is straightforward to define when a document *d satisfies* an atomic formula of the form $A = s$ or $B \sqsupseteq wp$, and then use this notion to define when *d* satisfies a query [11]. The example document given above satisfies the example query

4. THE NEURAL NETWORK CORPUS

The proof-of-concept corpus we use (called *NN corpus*) consists of a fraction of research papers from ResearchIndex [14,13] having Neural Networks as a subject. ResearchIndex, formerly known as Citeseer, is a digital library that targets the improvement in the dissemination of scientific literature. ResearchIndex indexes research articles in various formats and provides a variety of free services, such as full-text and citation indexing as well as paper statistics.

TABLE I. SOME CHARACTERISTICS OF THE NN CORPUS

Description	***Value***
Number of documents	10,426
Document vocabulary size	641,242
Maximum document size (words)	110,452
Minimum/maximum word size	1/35

The NN corpus consists of 10,426 scientific papers in English. Some important values for this corpus are summarised in Table I above. The documents were downloaded from the ResearchIndex site as postscript files and were converted to text files. Then all references and equations were removed and each word in the document was assigned a grammatical tag (e.g. noun, verb etc.) using a simple rule-based part of speech (POS) tagger [3]. This processing was necessary as a first step for the extraction of multi-word terms by the C-value/NC-value method described briefly in Section V.A and also in [9]. To use the corpus for our continuous query creation we also utilised the full citation graph of ResearchIndex.

Initially, we removed all the POS tags from all the documents. We then used the information from the full citation graph of ResearchIndex to extract the title, authors, abstract, and year of the publication. This information was not extracted from the actual corpus since the flat form of the documents contained considerable noise even after several rule-based filters were applied to it. The next step was to process the abstracts as POS-tagged text files, extracted from the original postscript files. After processing the abstracts we were able to identify the body of the document by excluding the information we already had in hand. When the processing phase was completed, we merged the different attributes extracted, along with the appropriate attribute tags. We then had at our disposal an attribute-tagged corpus with five fields: title, authors, abstract, body and year.

At this point we have to stress out that the information obtained from the citation graph was incomplete, resulting in documents without all the attribute fields filled in. This is actually not a problem in an experimental setting since in an information dissemination scenario users may post documents with only some of the attributes filled in. Table II gives some interesting measures of the fraction of documents out of the document corpus that contain each attribute, and summarises the fraction of documents that contain a specific number of attributes.

TABLE II. ATTRIBUTE STATISTICS

Attribute	***% fraction of documents***
title	63%
authors	58%
abstract	88%
body	86%
year	63%

Number of attributes	***% fraction of documents***
1	7.4%
2	28.0%
3	1.9%
4	16.0%
5	45.0%

5. CONTINUOUS QUERY GENERATION METHODOLOGY

The main construct in our profile creation process is that of a unit. Units in our context represent different entities that can be used to create a profile. The first two unit sets consist of proximity formulas created from multi-word terms, that were extracted from the NN corpus using the C-value/NC-value method described below. The third one is the set of all the nouns extracted from document abstracts, and the fourth one is the set of all author last names in the NN corpus documents. Combining units from these four sets in a well-defined way, allows us to create realistic profile databases in order to conduct our experiments.

A. Automatic Term Extraction

The multi-word terms used in the profiles for our experiments are extracted from the NN corpus using the C-value/NC-value approach of [9]. The process of identification of *terms* or *technical terms* or *terminological phrases* from a collection of documents belongs to the research area called *automatic term recognition*. The C- value/NC-value approach of [9] specifies the

"termhood" of a candidate multi-word term as the probability (co-location value) to be a real term. The C-value of a term is an enhancement of the common statistical measure of frequency of occurrence, incorporating information about nested terms, whereas NC-value embodies information form words that appear in the vicinity of terms in texts. Both methods have been shown to perform better than the classical frequency of occurrence measure in terms of precision and recall [8]. For details on the method the reader is invited to see [9, 8].

B. Creation of the Different Unit Sets

The creation of the first two unit sets was based on the extraction of multi-word terms from the corpus. To create these sets, a ranked list of multi-word terms was extracted from the corpus documents. We then, excluded from this list all terms that contained more than five words since they were noise produced by the C-/NC-value methods. Additionally, we specified an upper and lower NC-value cut-off threshold for the terms remaining in the list. These cut-off thresholds were used to increase the discriminating power of the set of terms. The upper cut-off threshold was used to exclude top ranked terms, that is terms that appear very often in corpus documents. Such an example is the 2-word term "neural networks" that is contained in most of our documents. Moreover the bottom ranked terms are also excluded from the list of the useful terms since they are mostly noise created from the procedure of transforming the original postscript files to simple text files. This processing resulted in a list containing 2-, 3-, 4- and 5-word terms, which was then used to create two different sets as follows.

Let $a_1a_2...an$, where each a_i is a word, be a multi-word term from the aforementioned list, containing *n* words. A proximity formula is created out of this term in the following two ways:

1. a1 $\prec_{[0,0]}$ a2 $\prec_{[0,0]}$ ... $\prec_{[0,0]}$ an. For each multi-word term in the list we introduce the proximity operator $\prec_{[0,0]}$ between the words of the multi-word term in order to create proximity formulas that represent strings. All the proximity formulas that are created this way form the first set of units named PF0, which stands for proximity formulas with word distance zero. The number of operands in these proximity formulas varies according to the number of words contained in the multi-word terms. The minimum number of words in a multi-word term is obviously two, whereas the maximum is five. An example of a unit in this set, which was produced from the term "inverse dynamic function", is inverse $\prec_{[0,0]}$ dynamic $\prec_{[0,0]}$ function

2. $a_1 \prec_{[0,k]} a_n$, where $1 \leq k \leq 10$. From each term in the list of multi-word terms we create proximity formulas with exactly two operands. These proximity formulas are created as follows. We replace *all* the middle words of the 3-, 4-, 5-word terms with the proximity operator $\prec_{[0,k]}$, specifying *k* to be a natural number drawn uniformly between 1 and 10. The choice of using a relatively small upper bound in the distance between two operands is inspired by the implementation of operator '*' and '*NEAR*' in Google and Yahoo! respectively. All the proximity formulas created this way form the second set of units named PF_k, since they are proximity formulas with word distance *k*. An example of a unit in this set could be *rbf* $\prec_{[0,6]}$ *networks*, which could be created from the term "rbf dynamic decay adjustment networks".

The second set of units used in the creation of our profiles database is the set of nouns that were extracted from document abstracts. The choice of nouns taken from document abstracts as

opposed to the whole document can be justified by the argument that the abstracts are expected to be a brief description of the work carried out in the paper thus, very appropriate to describe the content of a paper. The procedure of creating the set of nouns is as follows. First, we identified all the nouns in singular and plural form using the POS-tagged abstracts that were available to us. After that, we created a frequency-ranked list of these words and specified an upper/lower cut-off threshold to cut the most/least frequent words. The set of units that resulted from this procedure is denoted by NS, which stands for nouns.

The last set of units created is that of the authors' last names. We extracted all the names of the authors that were available to us from the corpus documents to obtain an author vocabulary V_{author} of 8,833 last names. Please notice that using this author vocabulary to uniformly draw author names for continuous queries is not a good choice, since authors that are more active or produce more important papers than others are expected to be used heavier in continuous queries. The criterion for identifying the more important authors is how many citations they get from papers written by other researchers. In the citation graph of the NN corpus this is captured by the in-degree of the papers as explained in [2]. The highest the in-degree for the papers of a specific author, the highest the probability for this author to appear in a profile. We define N_a to be the number of papers in the corpus that refer to at least one document of author a, and V_{author} the author vocabulary. N_a can easily be extracted from the full citation graph, and the author vocabulary is available to us from the NN corpus documents. Thus, the probability of author a to be used in a continuous query is:

$$P(a) = \frac{N_a}{\sum_{k \in V_{author}} N_k}$$

The above formula associates an author with the popularity of his writings and thus, with a probability of another researcher being interested in his work. To capture the probability distribution of the author surnames, we used a multi-set that contains an author surname Na times, and presents a power-law distribution (Figure 1). This can be explained by taking into account the general observation that in every scientific domain there exist a few heavily cited authors, while the rest receive less visibility (in terms of citations of their work). The unit multi-set described above is denoted by AS (author surnames).

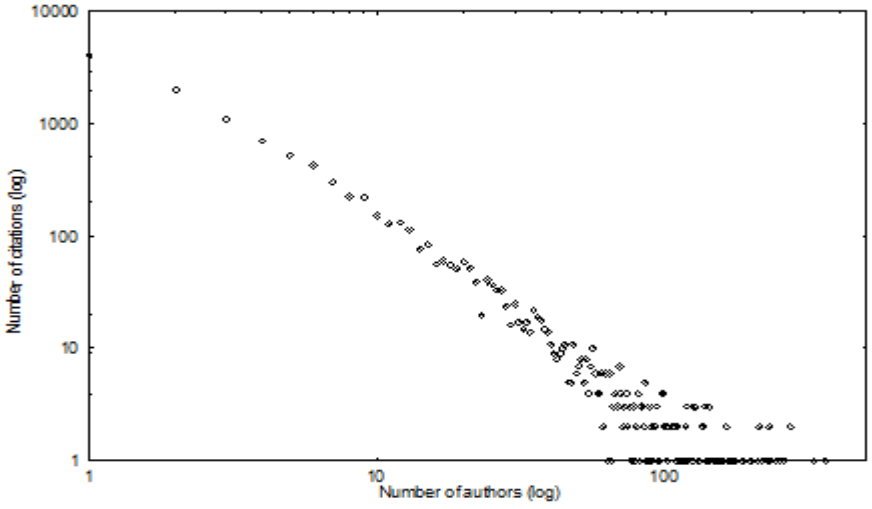

Figure 1. Distribution of citations among authors

C. Details

In this section, we provide details of how all the above extracted information is combined to create realistic continuous queries. A continuous query under the subset of query language AWP consists of a conjunction of atomic queries. These atomic queries can only be of the form A ⊒ wp, where A is an attribute and wp is a conjunction of words and proximity formulas. In the rest of this section we will examine the different types of atomic queries that can be created according to the attributes that are available to us.

TABLE III. SPECIFICS FOR THE CREATION OF ATOMIC FORMULAS

Attribute	***Participating unit sets***	***Indicatory value of σ***
title	PF0 / PFk / NS	0.4 / 0.4 /0.2
abstract	PF0 / PFk / NS	0.4 / 0.4 / 0.2
body	PF0 / PFk / NS	0.4 / 0.4 / 0.2
author	AS	1.0

In our context, creating a continuous query can be viewed as the problem of choosing with a probability distribution between units contained in different sets. Not all the sets of units participate in the creation of an atomic formula of a specific attribute. Moreover, different unit sets that participate in an atomic formula may have different selection probabilities (σ) in being chosen to participate in a profile. The unit sets that participate in the creation of an atomic formula, along with an indicatory value for the σ of each unit are summarised in Table III. Notice that these values may vary depending on the properties of the continuous query database to be generated.

In general, a creation of an atomic query is a 3-step process that can be described as follows:

1. Choose the number of units (or the *size* of an atomic query) *S*. This value is drawn uniformly from [1,S_{max}], where S_{max} is the maximum number of units in an atomic query. S_{max} is defined to be 2 for atomic formulas of author and title attributes, whereas it is set to 3 for the abstract and body attributes. This differentiation in S_{max} is due to the different number of words contained typically in the different attributes of a document.

2. Taking into account the units that may participate in a specific atomic formula, we pick *S* units from these sets according to the selection probabilities summarised in Table III.

3. Having chosen these units, we take their conjunction to create the atomic formula.

Thus, an atomic formula for the title attribute may be:

$$title \sqsupseteq (rbf \prec_{[0,6]} networks) \wedge java$$

which contains two units (remember that this is the maximum number of units allowed for the title attribute): unit ($rbf \prec_{[0,6]} networks$) drawn from unit set PF_k and unit *java* drawn from unit set

NS. Modifiying σ in the different unit sets results in controlling how often units of a specific set will appear in atomic queries of the corresponding attribute. Thus, other possible atomic formulas could be:

$$title \sqsupseteq implementation \wedge (dynamic \prec_{[0,0]} functions)$$
$$title \sqsupseteq real \prec_{[0,0]} world \prec_{[0,0]} application$$
$$title \sqsupseteq algorithm \wedge implementation$$

Atomic queries for abstract and body attributes are created in a similar way. The only differentiation between atomic formulas of different attributes is the value of σ for the unit sets and the maximum atomic query sizes.

At the same time creating atomic queries for attribute author is somewhat different since it may contain either one unit or a conjunction of two units from AS. Note that for the case of an atomic query for the author attribute using more words in conjunction would make the profile very specific, thus not suitable for an information alert setting. Note also, that proximity operations may also be used in these atomic queries (e.g., John $\prec$[0,0] Brown). However, the authors' first names were not available from the corpus documents so this option was not adopted. Some examples of such atomic queries are author $\sqsupseteq$ Brown or author $\sqsupseteq$ Smith $\wedge$ Johnson.

Finally, to decide which atomic queries will be introduced as conjuncts in each continuous query we assign selection probabilities to each one of the four types of atomic queries and according to this selection probabilities we include or exclude atomic queries. Each type of atomic query is (or is not) included in a profile independently of the rest of the types. For example, for a specific profile generation scenario if the selection probability of all four types of atomic queries is 85% then atomic queries for the author attribute will appear in the 85% of the profiles in the profile database. The same holds for the rest of the attribute types (title, abstract and body). At this point we should stress that in this way all possible combinations of atomic queries may appear in the generated continuous queries, and that a simple probability calculation allows us to control or exclude certain types of atomic queries.

6. CONCLUSIONS AND FUTURE WORK

In this work we presented a methodology for creating realistic artificial continuous query databases from any real-life (attribute-tagged) corpus, and as a proof-of-concept we applied it to the NN corpus. The robustness of the proposed methodology is highlighted not only by the publications in top-class venues that utilize it (e.g., [11, 12, 16, 18, 22]), but also by the different document corpora it was applied on (TREC *.gov, TREC ClueWeb09*, *OHSUMED*, *NN,* and others). Finally, interesting directions for future work include the design and implementation of modules for creating realistic vector space and semi-structured continuous queries.

ACKNOWLEDGEMENTS

The author would like to thank the anonymous reviewers for their comments and suggestions on improving the quality and presentation of the work

REFERENCES

[1] M. Altinel and M.J. Franklin. Efficient filtering of XML documents for selective dissemination of information. In VLDB, 2000.

[2] Y. An, J. Janssen, and E. Milios. Characterizing and Mining the Citation Graph of the Computer Science Literature. Technical Report CS-2001-02, Dalhousie University, Canada, 26 September 2001.

[3] Eric Brill. A simple rule-based part-of-speech tagger. In ANLP, 1992.

[4] C.-C. K. Chang, H. Garcia-Molina, and A. Paepcke. Predicate Rewriting for Translating Boolean Queries in a Heterogeneous Information System. ACM TODS, 17(1):1–39, 1999.

[5] R. Cover. The SGML/XML Web Page. http://www.oasis-open.org/cover/sgml-xml.html

[6] Y. Diao, M. Altinel, M.J. Franklin, H. Zhang, and P. Fischer. Path Sharing and Predicate Evaluation for High-Performance XML Filtering. ACM TODS, 2003.

[7] A.L. Diaz and D. Lovell. XML Generator. http://www.alphaworks.ibm.com/tech/xmlgenerator

[8] L. Dong. Automatic term extraction and similarity assessment in a domain specific document corpus. MSc thesis, 2002.

[9] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms:the C-value/NC-value method. IJDL, 5(2), 2000.

[10] T. J. Green, G. Miklau, M. Onizuka, and D. Suciu. Processing XML Streams with Deterministic Automata. In ICDT, 2003.

[11] M. Koubarakis, T. Koutris, C. Tryfonopoulos, and P. Raftopoulou. Information Alert in Distributed Digital Libraries: The Models, Languages and Architecture of DIAS. In ECDL, 2002.

[12] L. Zervakis, C. Tryfonopoulos, S. Skiadopoulos, and M. Koubarakis. Query Reorganisation Algorithms for Efficient Boolean Information Filtering. IEEE TKDE 29(2): 418-432, 2017.

[13] S. Lawrence, C. Lee Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. IEEE Computer, 32(6):67–71, 1999.

[14] ResearchIndex. The NEC Research Institute scientific literature digital library. http://www.researchindex.org

[15] I. Soboroff and S.E. Robertson. Building a filtering test collection for TREC 2002. In ACM SIGIR, 2003.

[16] C. Tryfonopoulos, M. Koubarakis, and Y. Drougas. Filtering Algorithms for Information Retrieval Models with Named Attributes and Proximity Operators. In ACM SIGIR, 2004.

[17] T.W. Yan and H. Garcia-Molina. Index structures for selective dissemination of information under the boolean model. ACM TODS, 19(2):332–364, 1994.

[18] C. Tryfonopoulos, M. Koubarakis, and Y. Drougas. Information filtering and query indexing for an information retrieval model. ACM TOIS, 2009.

[19] W. Rao, L. Chen, S. Chen, and S. Tarkoma. Evaluating continuous top-k queries over document streams. WWW, 2014.

[20] M. Sadoghi and H. Jacobsen. Analysis and optimization for boolean expression indexing. ACM TODS, 2013.

[21] http://dublincore.org/documents/dcmi-terms/

[22] L. Zervakis, C. Tryfonopoulos, S. Skiadopoulos, and M. Koubarakis. Query Reorganisation Algorithms for Efficient Boolean Information Filtering. In IEEE TKDE 29(2): 418-432, 2017.

INTENTIONAL BLANK

STUDY ON AUGMENTED CONTEXT INTERACTION SYSTEM FOR VIRTUAL REALITY ANIMATION USING WEARABLE TECHNOLOGY

Jiyoung Kang[1] and Jongkuk Lim[2]

[1]Graduate School of Cinematic Content, Dankook University, Korea
[2]Department of Computer Engineering, Dankook University, Korea

ABSTRACT

Currently, the virtual reality market is receiving global attention, and VR animation entertainment content is attracting attention as a new content market. In this study, we successfully develop these kinds of VR animations by focusing on interactions, the largest feature of virtual reality content, and we present a novel interaction method for virtual reality animations. We created ".FLY," a virtual reality movie, using wearable technology, and designed augmented context interactions to verify the emotional empathy and emotional immersion effect of the active story-based interactions experienced by the user in a virtual reality environment. The wearable bands used in the interactions were trained through programming, starting at the movie development stage. The audience viewed the interactive virtual reality movie and actively interacted through gesture interactions using the wearable bands.

KEYWORDS

VR, Interactive Animation, Augmented Context Interaction, Wearable Technology, Emotional Immersion

1. INTRODUCTION

Currently, as the world heads toward a fourth industrial revolution, there is considerable interest in the field of virtual reality. Virtual reality technology began in 1940 with flight simulators and has been continuously studied since then. In recent years, as wearable virtual reality devices such as HMDs (Head Mounted Displays) have been commercialized owing to advances in related hardware and software, virtual reality is garnering attention as the content platform of the future. However, the virtual reality market, which was expected to grow rapidly, has actually experienced faltering growth. This is because companies have not yet been able to establish a virtual reality ecosystem capable of expanding the virtual reality market due to the imbalances of a market developed with a focus on only hardware and related technology. In particular, the lack of content for a rapidly growing market centered on virtual reality devices is emerging as a problem in the virtual reality market. Because of this, the world is attempting all possible to dominate the virtual reality content market. In particular, games, which are rapidly converging

Dhinaharan Nagamalai et al. (Eds) : ITCSE, WiMo, ICAIT, ICDIPV, CRYPIS - 2018
pp. 47–58, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80705

with the virtual reality content market, are already expanding the market through specialized game distribution platforms such as Steam, and high-quality virtual reality games are being released. In addition, virtual reality movies are being actively developed and introduced at international conferences, movie festivals, and in Hollywood, including at SIGGRAPH, the Sundance Film Festival, and the Cannes Film Festival. In 2017 Google's "Pearl" became the first VR animation to be nominated for an Academy Award [1], and Penrose Studios' "Arden's Wake" was awarded Best VR Film at the 74th Venice Film Festival [2], showing the full potential of VR animation.

VR animation makes users emerge in a virtual space due to a combination of factors such as physical and task environment [3]. Physical environment produces a state of immersion and the task environment produces a state of involvement [4]. These environments can consist of combined hardware platform, software systems and interactive scenarios. There have been some researches about affective and cognitive elements, or endogenous factors, that contribute to an immersive experience and there is much emerging support within the literature for a relationship between presence and immersive experiences [5] [6] [7]. Researchers have also looked at a sense of flow, or deep involvement, with a virtual environment and immersion [8] [9] [10].

With the development of sensor technology, content that utilizes the motion of two hands or the whole body in real-time interaction with HMDs(Head Mounted Display) [11] while using wearable technology have recently appeared, yet it is difficult to find a case that utilizes real-time interaction in animation films. Therefore, this study proposes augmented context interaction using a wearable band in VR animation films based on interactive narratives. This study is concerned with the creation of immersive VR animations, and it analyzes the difference in storytelling between existing traditional movies and virtual reality movies with a focus on augmented context interaction. We aim to study the emotional effects of augmented context interaction in virtual reality movies that use wearable technology.

2. RESEARCH SCOPE AND METHODS

Virtual reality is a technology in which the users can have real-time interaction in a virtual space created by computer systems. It is a convergent technology that allows users to feel immersed in this virtual space through the five senses of the human body and that provides a feeling of presence as if the user actually existed in that space [12]. There are a variety of types of virtual reality content including CAVE type, desktop type, and third person type, but in this study we focus on wearable display HMD-based virtual reality movies, which are the most appropriate for personal media and growing at the fastest speed [13].

Virtual reality games targeted at existing game users are quickly becoming plentiful among HMD-based game content. However, movies, which are one of the most popular forms of entertainment, have not been able to form a virtual reality movie market as quickly as games. This is partly because HMD devices are associated with several problems such as cost and motion sickness from the cognitive dissonance caused by resolution and lag speed; however, the most important reason why audiences do not seek virtual reality movies is the lack of compelling content. Because the characteristics of virtual reality movies differ from those of existing movies, there is a need for movie storytelling based on deep research into these characteristics. In this study, we aim to analyze these virtual reality movie characteristics and propose an augmented context interaction method that uses wearable technology suited to these characteristics. In

addition, we perform experiments to verify the effect of the audience's emotional reaction on gesture interactions.

3. CURRENT VR FILM TECHNIQUE

3.1. Virtual Reality Movies and Presence

As with traditional 2D screen-based movies, virtual reality movies can largely be divided according to the methods by which they are made: live-action movies, graphics-based movies, and movies that were made by using a combination of these two methods. To date, research on these filmmaking methods made up most of the mainstream research on virtual reality movies, and research was not performed on methods of storytelling in virtual reality movies. This is because virtual reality movies are seen as a movie genre that is derived from the existing movie medium; thus, people have been unable to sense the need for a special grammar or type of storytelling that belongs only to virtual reality movies. However, when virtual reality movies borrow this kind of traditional movie storytelling style as-is it is difficult to properly show audiences the advantages held solely by virtual reality movies. Virtual reality movies that are produced with the same camera perspectives, etc. as the linear narratives of traditional movies can be called incomplete forms of virtual reality content that only add 360° imagery to traditional movies. However, we need a movie experience that is fundamentally different from existing traditional movies to enable audiences to feel the attraction of true virtual reality movies.

The ultimate goal of virtual reality is to allow the user to experience telepresence. Originally presence meant the sense of being within a certain environment, while telepresence can be called the experience of existing within a certain environment through a communications medium i.e., the mediated perception of an environment [14]. Feeling total presence through virtual reality movies is not possible simply through offering the point of view of an observer who turns their head around in a 360° space and sees a virtual space. That is, in order for the audience to feel this kind of presence in a virtual reality movie, we require communication beyond the passive communication of simply turning one's head. In short, we need mental and physical interactions between the audience and the medium called film. To do this, this paper proposes classifying this kind of communication as augmented context interaction.

3.2 Related Work

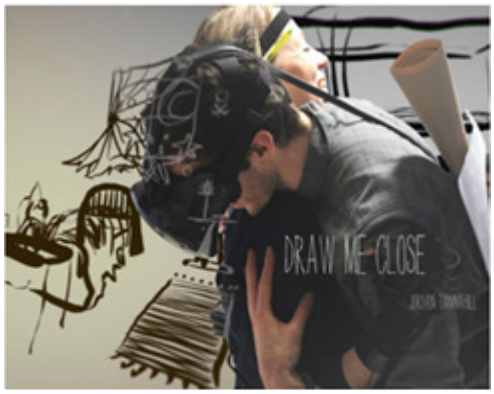

Figure 1: User interacting with an actor in the VR Film 'Draw Me Close'

Figure 2: Scene from the VR Film 'Draw Me Close'

In the spring of 2017, "Draw Me Close" [15] debuted at the Tribeca Film Festival, and it can be considered a typical example of active communication between an audience and a virtual reality movie (Fig. 1). The audience, equipped with an HMD and hand sensors, experiences 5-year-old Jordan's memories of his mother who passed away from cancer. People in the audience are equipped with an HMD and a motion capture system and experience something beyond direct participation in the space and time of the movie. The virtual space experienced through the HMD is shown in Fig.2, where we see Jordan's room and his mother expressed through line animation. In the virtual reality space, Jordan is lying down, and his mother pats him and hugs him as she puts him to sleep. There is active bodily interaction with the audience. Through this kind of embodied interaction, the audience feels immersion toward Jordan in the virtual space, and emotional storytelling takes place. However, in order to have this experience, the audience and the actors all have to wear special equipment to be motion captured, and there are problems with animation frame delays, etc. in the process of creating the real-time animation through the motion capture system. Moreover, a computer system with high specifications is needed to connect to the Vive HMD, and building the movie system is complicated.

Figure 3: Scene from VR Animation "Tree"

Active interaction in this kind of virtual reality animation can also be experienced in "Tree" [16], which debuted at the 2017 Sundance Film Festival and can be seen in Fig. 3. In this movie, members of the audience wear an Oculus Rift HMD and hold MIT-developed EMS (Electronic Muscle Stimulation) sensors in both hands and experience becoming a tree in the tropical rain forest. By bodily experiencing the whole cyclical process of the tree being born, growing up, and returning to the dirt, the audience has the mysterious experience of becoming part of nature. However, this movie does not provide the audience with emotional interactions that go beyond the intuitive experience of having the movement of their hands become the branches of the tree.

Conversely, due to the current limitations of technology, the disadvantages of such hand gesture inputs are: first, lack of comfort; second, every gesture can be interpreted by the system, whether or not it is intended, so the system must have well-defined means to detect the intention of the gesture; and finally, segmentation of hand gestures. In order to solve these, this study proposes augmented context interaction, which is suitable for VR animation stories experienced through wearable bands which the audience can easily put on. Through augmented context interaction, members of the audience, who easily put on the wearable bands without any special equipment and are in a free-hand state, can experience emotions and sympathy with the characters in a VR environment.

4. VIRTUAL REALITY ANIMATION USING WEARABLE TECHNOLOGY

4.1. Interactive VR Animation

Figure 4: Interactive VR Animation ".FLY"

In order to verify the emotional effects of these kinds of active communication in virtual spaces, this study created an interactive VR animation named ".FLY" [17]. ".FLY" is a virtual reality movie in which the audience interacts directly through gesture interactions using augmented content interaction, and it was shown by invitation in the Busan Film Festival's VR Theater in October 2017. Each member of the audience wore an HMD that incorporates a smartphone and wearable bands on the wrists that can recognize gestures to encounter a young girl within the movie. From a first person's view, the audience can either refuse the girl's hand as she asks for help or they can take her hand. In addition, they can console the girl when she collapses by patting her. Emotional storytelling occurs through these three different active bodily interactions.

4.2. Gesture Recognition Via Wearable Band

In order to recognize this kind of audience interaction in real-time as the movie plays, we created wearable bands that can be placed on people's wrists. According to the definition of the MIT Media Lab, a wearable device refers to anything that is attached to the body or worn and which can perform computing actions, including applications capable of performing some computing functions [18]. By placing wearable bands on the audience's wrists, we could immediately and naturally gather the movements made by their wrists as they watched the movie.

In order to recognize the audience's interactions as the movie plays in this way, we required the audience to put on wearable bands equipped with IMU (Inertial Measurement Unit) and watch the movie. IMU increases user convenience by combining a gyro and an accelerometer, and it contributes notably to raising the reliability and efficiency of the system [19]. Equipping the hand with a 6 degree-of-freedom sensor, we can map its position onto the position of a virtual hand in an immersive world and create the illusion that users can touch and grasp the girl in the VR animation using their own hands [20]. This allows users to interact naturally and intuitively in the immersive environment.

In order to provide gesture interactions to the audience in real-time as they watch the movie, it is necessary to follow a process of designing gesture interactions that are appropriate for the story and recognizing the designed interactions during the planning stage of the movie. To do this, we first gathered acceleration and gyroscope sensor data from the bands worn on the wrists, as in Fig. 5. Afterward, we selected features from the collected sensor data from which to distinguish actions. For example, distinguishing two kinds of actions requires us to calculate the speed, continuous motion values, etc. of the action as features. Subsequently, we implemented a methodology for distinguishing actions from the corresponding features through programming.

We implemented a rule-based gesture recognition system. Our interaction system design has the simplest form to minimize a computation burden. Accelerometer and gyroscope sensor are used in this system. First, we extract features as following, x, y, z, magnitude, minV, maxV. First three features, x, y, and z, are raw data from the sensor. Magnitude is defined as $\sqrt{x^2 + y^2 + z^2}$. Lastly, mixV and maxV are computed with each axis of sensor data. It is defined as $minV = [minX, minY, minZ]$, $maxV = [maxX, maxY, maxZ]$ and $minX = \min(minX, x_t).t$ is time and rest features are equivalently computed as minX. However, minV and maxV are initialized to 0 when the user attempts to start gesture interaction to avoid unintentional gesture. Features are computed with each sensor separately which generates 12 features. Every gesture can be divided into two steps First it waits until magnitude value exceeds threshold to recognize whether user has intention to make a gesture. Then it performs the actual recognition process. Recognizing holding a hand gesture exploits x, y, z values of accelerometer. It is watched if user holds the pre-defined angle for 3 seconds. Stroking hair exploits minV and maxV values of accelerometer and gyroscope. Thresholds of minimum and maximum are used to filter when the user tries to transit hand movement direction. Refusing a hand gesture can be thought of as a sequence of actions followed by holding and withdrawing a hand.

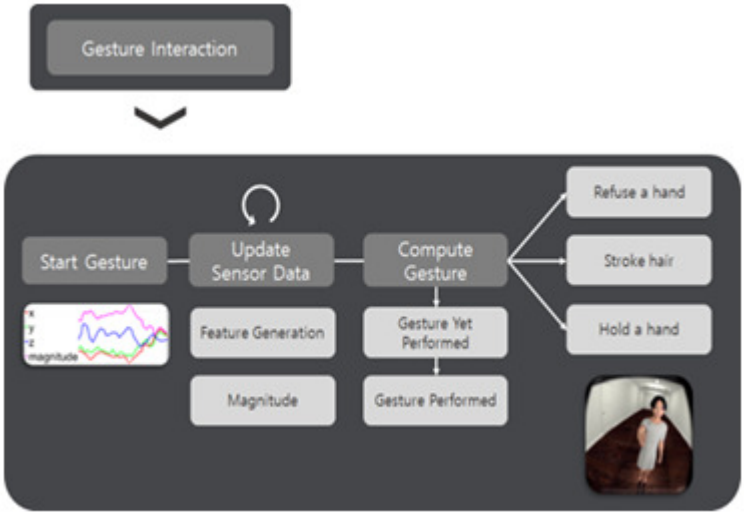

Figure 5: Process of gesture recognition

4.3. Augmented Context Interaction

Along with the recent growth of virtual reality content, continuing research has been focused on the interaction between users and content in order to provide the user with experiences optimized for interactions with content. In this research, there is interest in the extent to which the user can

use the content in a natural way, and this is creating a turning point in the development of natural user interface (NUI) technology. In the most typical method of NUI, a voice is played, and currently this is used in many products and services. Along with voice, gestures are an often-used user interface. Gestures include not only the intentional actions a user undertakes to convey their meaning, but also the unconscious, meaningless actions that they make.

In gesture recognition technology, sensors or equipment can be broadly divided into the contact type, in which data are acquired as the user directly makes contact via their body, and the non-contact type, in which data are acquired using long-distance and short-distance sensors. In this study we used the contact type of gesture recognition, which utilizes wearable technology worn on the user's wrists. This is because this equipment is very well suited to virtual reality environments because it can detect detailed gestures by simply being worn, without any separate camera equipment or limits on the user's area of activity or space.

When designing the interactions for this virtual reality movie, we used a concept known as augmented context interaction to move the audience's hearts and naturally induce interactions from the audience. This means augmenting the context information contained in the story, characters, or scenarios of the movie for an interaction given to the audience within the movie such that the interaction can be performed without resistance. That is, augmented context interaction is provided such that interactions can be performed naturally such as when the audience naturally reaches out as the girl, who is in danger, reaches out and asks for help or when they stroke her hair after she collapses on the cold street.

The first definition of context-aware applications, provided by Schilit and Theimer [21], ranged from applications that are simply informed about context to applications that adapt themselves to context. Context-awareness has been described in various research efforts with the following language: adaptive [22], reactive [23], responsive [24], situated [25], context-sensitive [26], and environment directed [27]. Previous research also described context-aware computing in two main ways: using context and adapting to context. In this study we used adapting to context and proposed gesture interactions suitable to scenarios in which there is interaction with the story and characters of the animation. This enables the audience to have augmented bodily interactions with characters that were limited in the virtual reality animation environment. Even though these interactions are not felt directly through the skin as in the real world, the act of stroking the character in the virtual space with one's own hand or holding her hand is directly experienced; thus, it is possible to allow the audience to feel more immersion and emotional storytelling.

Performing this kind of augmented context interaction design requires us to make it possible for the audience to naturally recognize what kind of gestures they must perform based on their past experiences. Ulmer and Ishii call this "Expressive Representation (ER)," which is when the user already reads and interprets the circumstances during the interaction and devises behaviors and modifications [28]. That is, the audience understands and analyzes the situation within the movie through this kind of active interaction, enabling meaning to be created by each user.

Along with this, the movie introduces the audiences to virtual reality, makes a life exist within the movie, and causes direct bodily interactions. This is because the VR environment is the optimal environment for tracking the audience's potential movements, and it can provide richer interactions than other media [29]. According to previous research, the user has a deeper immersion effect toward content when performing bodily interactions in a virtual space than when viewing video images [30]. This means that if suitable bodily interactions are used in

harmony with the story in a virtual reality movie as in the examples above, the audience's emotional immersion is raised, and a greater degree of immersion occurs.

Figure 6: Girl reaching out to the viewer

Along with this, the movie introduces the audiences to virtual reality, makes a life exist within the movie, and causes direct bodily interactions. This is because the VR environment is the optimal environment for tracking the audience's potential movements, and it can provide richer interactions than other media [29]. According to previous research, the user has a deeper immersion effect toward content when performing bodily interactions in a virtual space than when viewing video images [30]. This means that if suitable bodily interactions are used in harmony with the story in a virtual reality movie as in the examples above, the audience's emotional immersion is raised, and a greater degree of immersion occurs.

5. RESULTS

We conducted experiments to confirm that the active communication achieved through augmented context interaction caused emotional empathy by directly inviting the audience into the virtual reality space and providing an embodied experience. The users included 20 men and women between the ages of 20 and 40. In the tests, the users were divided into two groups of 10 people: user group A wore the wearable bands and had active interactions, and user group B watched the VR movie and did not have active interactions. After viewings, we evaluated the users' emotional immersion and presence through a survey.

Question 1 measures the emotion of sympathy felt by the audience when experiencing empathy toward the story as they watched the movie and encountered the girl from the main character's 1st-person point of view. Through this question, we aimed to evaluate the audience's emotional empathy effect. Questions 2 evaluated whether the audience felt as though they existed within the virtual space and the degree of presence they felt interacting with the girl character in the movie. We used these questions with the aim of evaluating the immersion effect.

1) If you felt sympathy or pity toward the girl as you watched the movie, to what extent did you have these feelings? (5 is the highest level of sympathy)

2) How much immersion did you feel toward the main character in the movie? (5 is the highest level of immersion)

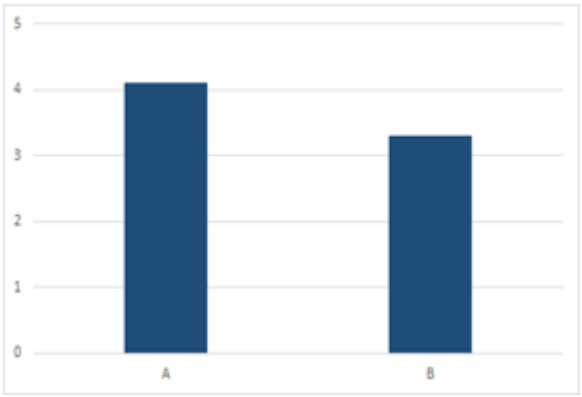

Figure 7: Result for empathy

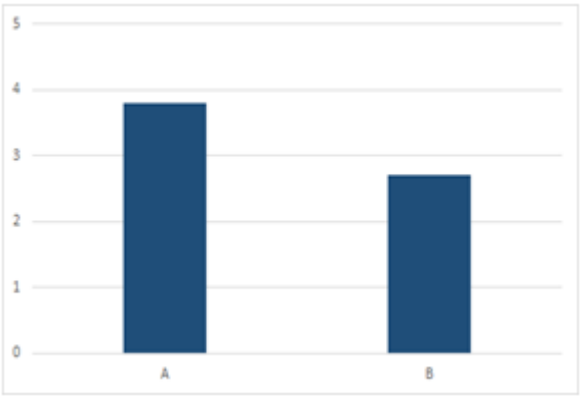

Figure 8: Result for immersion

Through the above survey we found that user group A, which wore the wearable bands and experienced active interaction, had more sympathetic emotions toward the girl in the movie at a score of 4.1 than user group B which did not experience gesture interactions and had a score of 3.3. Furthermore, group A's immersion as the main character of the movie was higher at 3.8 than group B's, which was 2.7. It can be seen that the users of group A, who wore the wearable bands and directly performed gesture interactions according to the given situation in the virtual reality environment, felt considerable emotional empathy and a sense of immersion compared to when gesture interactions were not performed during viewing. This means that when the audience is immersed in a virtual reality environment and views a movie, if augmented context interaction is properly used in a way that is suitable for the story and scenarios of the movie, it can strengthen the immersion and emotional storytelling of the virtual reality content.

6. CONCLUSIONS

We described research aimed at developing a method for virtual reality animation storytelling that combines a form of active communication known as augmented context interactions with virtual reality animation. We provided audiences with augmented context interaction by using wearable bands to provide gesture interactions that are emotionally and bodily suitable for the story and scenarios of a virtual reality movie. We also conducted experiments to confirm that the active

communication achieved through augmented context interaction caused emotional empathy and immersion.

For the next step, we will perform user tests to verify that the degree of emotional immersion in an audience experiencing augmented context interactions compare to passive traditional interactions. Also, in this study, we performed research on augmented context interaction in virtual reality movies, and we must perform a variety of research on stories and dramatic methods suitable for this kind of interaction. Many tasks remain in the research and development of technology and content related to virtual reality movies to enable this kind of virtual reality animation to take a place in the mainstream cultural content industry

ACKNOWLEDGEMENTS

This research was conducted with the support of the 2017 National Research Foundation of Korea (2017R1D1A1B03027954), and we would like to extend our appreciation to the relevant government departments.

REFERENCES

[1] POLYGON. (2017). Watch VR's first Oscar-nominated short film. https://www.polygon.com/2017/1/24/14370892/virtual-reality-first-oscar-nominated-short-film-pearl

[2] CHINA GLOBAL TELEVISION NETWORK. (2017). Animated film "Arden's Wake" wins best virtual reality award in Venice. https://news.cgtn.com/news/326b444d34557a6333566d54/share_p.html

[3] Slater, M., & Usoh, M. (1993). Presence in Immersive Virtual Environments. IEEE Virtual Reality Annual International Symposium, 90-96.

[4] Chertoff, Dustin B., Brian Goldiez, and Joseph J. LaViola. (2010). "Virtual Experience Test: A virtual environment evaluation questionnaire." In Virtual Reality Conference (VR), 2010 IEEE, 103-110.

[5] Chertoff, D. B., Schatz, S. L., McDaniel, R., & Bowers, C. A. (2008). Improving presence theory through experiential design.Presence: Teleoperators and Virtual Environments , 17 (4), 405-413.

[6] Davide, F., & Walker, R. (2003). Engineering Presence: an Experimental Strategy. In G. Riva, F. Davide, & W. A. IJSselsteijn, Being There: Concepts, effects and measurement of user presence in synthetic environments. Amsterdam, The Netherlands: Ios Press. 41-57.

[7] IJsselsteijn, W. A. (2003). Presence in the Past: what can we learn from Media History? In G. Riva, F. Davide, & W. A. Ijsselsteijn, Being There: Concepts, effects and measurement of user presence in synthetic environments. Amsterdam, The Netherlands: Ios Press. 17-40.

[8] Chen, X. (2006). Flow in Games. University of South California.

[9] Fu, F.-L., Su, R.-C., & Yu, S.-C. (2009). EGameFlow: A scale to measure learner's enjoyment of e-learning games. Computers and Education , 52, 101-112.

[10] Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. Computers in Entertainment, 3 (3).

[11] Sutherland IE. (1968). A head-mounted three dimensional display. InProceedings of the December 9-11, fall joint computer conference, part I 1968 Dec 9. 757-764. ACM.

[12] Kim, Y. K., Yoon, Y. S., Oh, T. G., Hwangbo, Y. H., and Hwang, J. H. (2017). Real-time VR strategy chess game using motion recognition. J. Digital Contents Soc. 18, 2.

[13] Ullmer, B., and Ishii, H. (2000). Emerging frameworks for tangible user interfaces. IBM Syst. J. 39, 915-931.

[14] Gunawardena, C. N., & Zittle, F. J. (1997). Social presence as a predictor of satisfaction within a computer-mediated conferencing environment. American journal of distance education, 11(3), 8-26.

[15] TRIBECA FILM FESTIVAL. (2017). Draw me close. https://tribecafilm.com/filmguide/draw-me-close-2017

[16] TREE OFFICIAL. (2017). The film. https://www.treeofficial.com/

[17] BUSAN INTERNATIONAL FILM FESTIVAL. (2017). 22nd International Film Festival, 12-21 October, 2017. http://www.biff.kr/

[18] MIT Media Lab. (2017). Relational AI. http://www.media.mit.edu

[19] Jacoby, R., Ferneau, M., Humphries, J. (1994). Gestural Interaction in a Virtual Environment. Stereoscopic Displays and Virtual Reality Systems, SPIE 2177, 355-364

[20] Poupyrev, I., Billinghurst, M., Weghorst, S. and Ichikawa, T. (1996), November. The go-go interaction technique: non-linear mapping for direct manipulation in VR. In Proceedings of the 9th annual ACM symposium on User interface software and technology. 79-80. ACM.

[21] Schilit, B. and Theimer, M. (1994). Disseminating active map information to mobile hosts. IEEE Netw. 8, 22–32.

[22] Brown, M. G. (1996). Supporting user mobility. Ed. by J. L. Encarnação and J. M. Rabaey. Mobile Communications, 69–77. Springer, Boston.

[23] Cooperstock, J. R., Tanikoshi, K., Beirne, G., Narine, T., and Buxton, W. A. (1995). Evolution of a reactive environment. In Proceedings of the SIGCHI Conference on Human Factors In Computing Systems, 170–177.

[24] Elrod, S., Hall, G., Costanza, R., Dixon, M., and Des Rivieres, J. (1993). Responsive office environments. Commun. ACM, 36, 84–85.

[25] Hull, R., Neaves, P., and Bedford-Roberts, J. (1997). Towards situated computing. In First International Symposium on Wearable Computers, 146–153.

[26] Rekimoto, J., Ayatsuka, Y., Hayashi, K. (1998). Augment-able Reality: Situated Communication through Physical and Digital Spaces. 2nd International Symposium on Wearable Computers. 68-75

[27] Fickas, S., Korteum, G., Segall, Z. (1997). Software Organization for Dynamic and Adaptable Wearable Systems. 1st International Symposium on Wearable Computers, 56-63

[28] Steuer, J. 1992. Defining Virtual Reality: Dimensions Determining Telepresence. J. Commun. 42, 73-93.

[29] Lanier, J. 2001. Virtually there. Scientific American, 66-75.

[30] Bailenson, J., Patel, K., Nielsen, A., Bajscy, R., Jung, S.H., and Kurillo, G. 2008. The effect of interactivity on learning physical actions in virtual reality. Media Psych. 11, 354-376.

AUTHORS

Jiyoung Kang
Associate Professor at Dankook University
Ph.D, Korea Advanced Institute of Science and Technology
M.P.S, New York University

Jongkuk Lim
Ph.D Candidate
Department of Computer Engineering,
Dankook University

TOWARDS AN ASSESSMENT OF CLOUD E-HEALTH PROJECT RISK: AN EMPIRICAL STUDY

Bouchaib Bahli

School of Information Technology Management, Ted Rogers School of Management, Ryerson University, Toronto, Canada

ABSTRACT

The introduction of information technology and telecommunications (ITC) in the health care sector has brought so many benefits to the health operators, managers, and patients. However, the increasing use and the application of ITC to the management and delivery of health care well known as e-health has been associated with several e-health risks that need to be examined. In this paper we point out several shortcomings of current risk conceptualization and operationalization, particularly they do not address the integration of a variety of risk components, which are crucial for capturing the essence of e-health risks. To fill this gap and drawing on risk analysis perspective we present and discuss a formal framework for e-health cloud computing project risks that captures potential scenarios, their likelihood and, the associated negative consequences. E-health risks were identified in the literature and a cluster analysis was used to classify different risks into several risk domains according to the developed e-health risk framework. Results show several domains including privacy, security, safety, liability, operational, project and business e-health risks. Implications for researchers and managers are also discussed.

KEYWORDS

e-health, e-health risks, risk assessment, health care management

1. INTRODUCTION

This document describes, and is written to conform to, author guidelines for the journals of AIRCC series. It is prepared in Microsoft Word as a .doc document. Although other means of preparation are acceptable, final, camera-ready versions must conform to this layout. Microsoft Word terminology is used where appropriate in this document. Although formatting instructions may often appear daunting, the simplest approach is to use this template and insert headings and text into it as appropriate.

The use of technology to assist in the delivery of quality patient care covers a vast areas from biomechanical devices to robotics to the electronic medical record to email. More particularly, Internet as a source of health information and connectivity between providers and payers has increased interest in e-health as a channel for the delivery of health-related products and services (Trudel et al; 2012). Of the 137 million Americans who surf the Internet, more than 60 percent use the resource for health advice (Harris, 2002). In addition, 90% of adults would like to be able to communicate with their physicians on line and the number of adults who have looked for health information has climbed from 54 million to 110 million (Harris, 2002). In the US alone the

Dhinaharan Nagamalai et al. (Eds) : ITCSE, WiMo, ICAIT, ICDIPV, CRYPIS - 2018
pp. 59–66, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80706

move to a fully integrated e-Health system has been estimated to improve efficiency and reduce costs, saving some $81bn (Appari & Johnson, 2009). Sometimes, the term e-health has been used very loosely to include any electronic healthcare-related activity (DeLuca and Enmark, 2000). One widely definition of e-health is the one adopted by HIMSS' e-Health SIG (2003). E-health is defined as the application of Internet and other related technologies in the healthcare industry to improve the access, efficiency, effectiveness, and quality of clinical and business processes utilized by healthcare organizations, practitioners, patients, and consumers to improve the health status of patients.

The potential benefits of computerization are considerable. E-health systems can facilitate access to patients medical records, improve the quality of care and the accuracy of treatment decisions, achieve cost savings, and promote clinical research (Baron, 2005) and some health care providers with e-health systems already report better outcomes, fewer complications, lower costs, and fewer malpractice claim payments (Amarasingham et al. 2009). Without discounting any of these potential benefits, e-health systems continue to face challenges and associated risks within the health industry (Tsiknakis and Kouroubali 2009).

This paper focuses on the risks associated with e-health systems and on concerns associated with their use. We argue that despite the promise of this technology, the implementation of e-health systems must proceed with both caution and appropriate oversight. E-health systems give rise to new risks for health care providers and patients alike. Computerized information is vulnerable to large-scale privacy violations associated with hacking, computer theft, malicious electronic distribution, or accidental disclosure, such as sending a file to the wrong e-mail address. Once data security is breached, the most private information can be dispersed on the Internet to a worldwide audience (Hoffman and Podgurski, 2009). Disclosure of psychiatric or sexual histories or other sensitive information can, among other harms, lead to profound embarrassment, ruined careers, or loss of professional and personal opportunities. These, in turn, can generate litigation against those responsible for security breaches (Hoffman and Podgurski, 2009).

This study's contributions to research and practice are twofold. First, we point out several shortcomings of current risk conceptualization and operationalization and present a more comprehensive approach to risk assessment. Second, we present and discuss e-health risks and map them into the developed risk framework. Our framework provides a step further in systematically assessing e-health risks and provides health operators and managers with a tool that captures a variety of potential scenarios, their likelihood and, the associated negative consequences. Further, we discuss both research and practical implications of this framework, research limitations and further research. The rest of the paper is organized as follows. First, we provide the theoretical foundation for the study. Second, we present the research developed e-health risk framework. Next, we discuss the implications of this framework for research and practice, study's limitations and directions for further research.

2. E-HEALTH RISK: A CONCEPTUALIZATION

For many health care organizations that are providing health services to patients, doctors, nurses, health operators, etc, one real concern is that of the risks associated with conducting transactions through electronic channels (Baker et al., 2005). While the risk construct itself has been conceptualized and measured across disciplines (Bahli and Rivard, 2003), the e-health literature suggests that the conceptualization and measurement of e-health risks is yet to be examined in a formal manner. The first objective of the present research is to refine and validate the conceptualization and measurement of e-health risk in a comprehensive research framework embedding the different types and categories of e-health risks. The term risk is likely to be one of

the most often used words in modern language. Every day, in extremely different circumstances, people use the term risk, be it to talk about the probability of a snowfall or the variability of their investments (Bahli and Rivard, 2003). Risk is defined along a decision theoretic view as the likely variability of future returns from an asset, equity or investment. The behavioural perspective associates risk with the magnitude of a negative consequence of a decision. In this view, a risky choice is one that contains a threat of poor performance. In information systems, the notion of risk exposure – that is, the combination of the probability of occurrence of an undesirable event and the amount of loss related to this event – is often used (Boehm, 1991; Barki et al., 2001).

In their widely cited paper entitled "On the Quantitative Definition of Risk," Kaplan and Garrick (1981) criticized the fact that researchers often took into account the sole probability of occurrence of an undesirable event in defining risk. Furthermore, the traditional expected consequence representation of risk (generally referred to as risk exposure) is deemed inappropriate by Kaplan and Garrick since this representation assumes a risk-neutral decision maker. According to these authors, most people would rather judge a low-probability-high-consequence scenario as more undesirable than a high-probability-low consequence scenario even if the expected consequences of the two events were equal. In other words, this means that concepts like frequency-severity diagrams have the undesirable property that very different situations, among which a rational risk adverse decision maker might have clear preferences, could be mapped into identical diagrams. Rather, Kaplan and Garrick argued that three questions ought to be addressed in order to assess risk. These questions are: what can happen? (i.e., What can go wrong?), how likely is it that will happen?, if it does happen, what are the consequences?. Kaplan and Garrick proposed a general definition of risk as a complete set of triplets involving scenarios (what can happen?), the likelihood of each scenario (how is it likely to happen?), and the consequences or evaluation measure of each scenario, that is, the measure of damage. "To answer these questions we would make a list of outcomes or "scenarios" as suggested in Table 1. The ith line in Table 1 can be thought of as a triplet:

< pi, si, xi>

Where si is the scenario, pi is the probability of that scenario, xi is the consequence

Table 1. Scenario List (Kaplan and Garrick, 1981)

Likelihood	Scenario	Consequence
p1	s1	x1
pn	s2	x2
.	.	.
.	.	.
.	.	.
pn	sn	xn

This conceptualization of risk seems to be more appropriate for the present study in two ways: First, it allows capturing several scenarios that may emerge due to the use and application of e-health systems, the likelihood of these scenarios happen and, the associated negative consequences. Second, managers can visualize a series of triplets and decide which scenarios need to be avoided or attenuated and select appropriate measures to mitigate them. The following section presents an application of Garrick and Kaplan's risk framework to capture e-health risks. A cluster analysis was used on a variety of e-health risks in the literature. The results were then mapped into Garrick and Kaplan's risk framework.

3. DESIGN AND METHOD

The proposed research consists of surveying 10 IT managers in the health sector in Croatia. These managers had no obligation to respond to our questions and they were not evaluated to do so, hence, reducing research bias. The average years of experience in the IT field was 8.65. The managers come from the health sector: Pharmaceutical (6), biotechnology industry (2), hospitals (2). The selected 10 managers were chosen because of accessibility opportunity and it took four seminars to identify them. Our choice was partly opportunistic, in that these managers attended a seminar on IT risks, thus making accessibility less of a problem. More importantly, these managers met our criteria of suitable cases on e-health projects. These managers were involved at different degrees in e-health projects of their organizations.

After explaining the concept of risk and its components as suggested by Kaplan and Garrick (1981), we asked the respondents about their perceptions on the potential e-health project risks. For instance, if one manager perceived a certain type of e-health risks he or she needs to explain what constitutes their judgment. All responses were transcribed, interpreted and analyzed. The following section describes the data analysis process. We limit our study to the mapping of all e-health risks into the RISC framework.

Categorical analysis attempts to make valid inferences from studied texts to their underlying meaning in terms of pre-specified set of categories (Weber 1985). The goal for using categorical analysis was to develop a systematic representation of the different categories of e-health project risks and thereby to reveal their varying foci and rationale. As recommended by Glaser and Straus (Glaser and Straus, 1967), we have produced some explanations of theoretical concepts and patterns (Orlikowski, 1993) of each of the risk categories. Three steps are required: (1) initial analysis of transcripts where the responses were transcribed and analyzed. We highlighted comments about managers' perceived e-health project risks. (2) Interpretation of transcripts to dissect patterns in responses. (3) Analysis of the interpretations. Two people performed this task individually, then, we corroborate both analyses as suggested by Tesch (1990) in order to decontextualize comments. Both individuals agree on the labeling of each level. The coding was conducted by classifying every risk factors, scenario and consequences using RISC components. Before actual coding, we agreed on a number of coding rules. Each risk component was assigned to one RISC coding scheme. The coding was based on component content description. Sometimes this lead to a further reading of the component description in the main body of the text to further clarify its meaning. In our situation, this analysis technique helps clarify the e-health risk categories. In particular, categorical analysis reveals several groups or risk domains. Then, within each domain, we identified scenarios, risk factors and the associated consequences. Several risk domains were identified: privacy risks, security risks, litigation and liability risks, safety risks, project risks and, operational risks.

Privacy Risks include the unauthorized collection, use and disclosure of personal health information and, any threat to the ability of the patient to exercise any right under privacy legislation (Karsh et al. 2006). Computerized information is vulnerable to large-scale privacy violations associated with hacking, computer theft, malicious electronic distribution, or accidental disclosure. Once data security is breached, the most private information can be dispersed on the Internet to a worldwide audience (Hoffman et Podgurski , 2009).

Security Risks are breaches of confidentiality, integrity and availability of personal health information and/or critical health information systems (Kolkowska, Hedstrom, & Karlsson, 2009). Security risks include the loss, corruption or unauthorized modification of personal health information and; loss of critical ICT services (Baker et al., 2005). Security of data in medical applications is particular complex because patient data is typically fragmented controlled by

whoever provided health services. Moreover, the security mechanism must be arranged so that users can quickly share information in the event of an emergency (Lorence and Churchil, 2005). On the other hand many healthcare professional are reluctant to use information and communication tools due to the security risks entailed.

Safety risks include physical or mental harm to patients and health care providers (including death). Safety Risks. Patient injury or death, health provider injury or death and Population injury or death (Karsh et al., 2006). Security issues are the most likely sources of e-health safety problems. Integrity and availability issues could certainly impact patient and health provider safety (Lohr et al. 2009), particularly as we become more dependent on e-health programs. Consider what could happen to a patient if a security breach brought down an e-health portal that provided access to critical health information systems or a software bug that causes a lab system to deliver inaccurate test results (Balka et al. 2006). Human factors issues – where the human/information system interface fails. This could include user interfaces that are confusing, overly complex procedures that promote error or failure to catch common user errors (e.g. input procedures that make it easy to enter the wrong data or displays that make it easy to misinterpret data).

Patients who learn that their medical information has been inappropriately disclosed to third parties may be inclined to sue their physicians. In fact, patients might initiate litigation not only when the physician has carelessly or intentionally disclosed private information, but also when the disclosure occurred because of hacking or a system defect (Hoffman and Podgursk, 2009). Use of e-health systems could generate negligence claims against providers for a variety of reasons. EHR system operation can be time-consuming and burdensome, and increased work demands could cause rushed physicians to make medical mistakes. Greater access to existing diagnostic data and economic pressures to avoid duplicating tests could lead to errors from inappropriate reliance on outdated or inadequate prior testing. Mistakes may also result from data entry errors (Hoffman and Podgursk, 2009).

Project risks include cost overruns, scope creep, unacceptable delays, failure to deliver required functionality or project failure (Balka et al., 2006). Heeks (2006) estimated that 20-25 percent of IT projects in healthcare are total failures and 33-60 percent are partially successful. This is attributed to poor IT investment decision making and to the increasing complexity of IT implementations in recent years (Trudel, Paré & Laflamme, 2012). Prior research shows that the management of IS projects is often marked by inadequate planning, a poor grasp of the overall development process, and no clear management framework, even as the focus shifts from a technology perspective to a more process-centric view (Agarwal and Rathod, 2006). .

Operational risks include incompatible technology, obsolescence, inability to meet service levels, lack of skilled human resources (Barki et al. 2001). While an e-health project will end, the e-health program will extend through time until the program is eventually replaced or retired. The operations phase is the phase in which identified benefits will materialize. During this phase threats to the continuing success of the program may materialize (Schabetsberger et al. 2006). There is a close relationship between operational risks and security risks. Adapting the security threat and risk assessment methodology to address broader operational risks would address many of the issues arising in the operational environment (Garg et al, 2005).

Table 2. E-health Risk Framework

Risk Type	Risk Factors	Scenarios	Consequences
Privacy Risks	Weak patient consent procedures	Unauthorized collection Unauthorized use Unauthorized disclosure Denial patient rights	Loss damage of the patient privacy
Security Risks	Malicious use or attack Natural IT service failure	Loss of personal health information Unauthorized modification of health information Loss of critical IT services	Loss damage of the patient security
Saftey Risks	Critical failure of IT services Failure to communicate critical safety procedures	Patient injury or death Health provider injury or death Population injury or death	Loss damage of the patient safety
Liability Risks	Litigation Political constraints User resistance Loss of critical resources	Legal liability Financial loss Business interruption External interefernces Rejection by users	Loss damage of the business
Project Risks	Unrealistic expectations Resources not available Scope creep Poor budgeting	Scope compromise Time delay Cost overrun	Loss damage of the project failure
Operational Risks	Poor maintenance Systems obsolescence Lack of skilled resources	Loss of service	Loss damage of the operational business

4. PRACTICAL IMPLICATIONS AND EXPECTED CONTRIBUTIONS

The proposed research paper contributes to a better understanding of how e-Health risk is defined and measured. It also provides a formal tool for the assessment of e-Health risks. The identification of the components of e-Health risk is based on Seaton (2007) categorization of risks. We mapped this categorization into the risk framework developed by Garrick and Kaplan (1981). The proposed research will contribute to both research and practice. For research, the main contribution is to further our understanding of how e-health risk components are modeled. In addition, this research could contribute to the field of e-health by providing a better understanding of risk in general and e-health risks in particular. For practice, understanding the polymorphism of risk could help healthcare organizations more adequately and fully exploit the benefits of e-health systems. From this research it can be seen that while e-health systems are bringing many benefits to the work flow and practices of these health care practitioners, there is a need for further development of the technology in order for it to fulfil its potential and truly achieve a categorisation of success as indicated by user satisfaction with e-health systems.

REFERENCES

[1] Agarwal, N. and Rathod, U. 2006. Defining "success" for software projects: An exploratory revelation. International Journal of Project Management, 24, pp. 358-370.

[2] Appari, A. and Johnson, E. 2010. Information security and privacy in healthcare: current state of research". Int. J. of Internet and Enterprise Management, 6, 4, pp. 279-314.

[3] Amarasingham, R et al. 2009. Clinical Information Technologies and Inpatient Outcomes: A Multiple Hospital Study. Archives of Internal Medicine, 169, 2, pp. 108-114

[4] Baron, J. et al. 2005. Electronic Health Records: Just around the Corner? Or over the Cliff?. Annals of Internal Medicine, 143, 3, pp. 222-226.

[5] Bahli, B. and Rivard, S. 2003. The information technology outsourcing risk: a transaction cost and agency theory-based perspective. Journal of Information Technology, 18, pp. 211-221

[6] Baker et al. 2005. The Canadian adverse events study: the incidence of adverse events among hospital patients in Canada, CMAJ, 170, 11, pp. 1684-1689.

[7] Balka et al. 2006. Technology, governance and patient safety: Systems issues in technology and patient safety, International Journal of Medical Informatics, 76, pp. 35-37.

[8] Boehm, B.W. 1991. Software risk management: principles and practices. IEEE Software, 12, pp. 32-41.

[9] Currie, W and Finnegan, D. 2011. The Policy- Practice Nexus of Electronic Health Records Adoption in the UK NHS: An Institutional Analysis. Journal of Enterprise Information Management, 24, 2, pp. 146-170.

[10] Currie, W and Guah, W.M. 2007. A national program for IT in the organisational field of healthcare: A Example of Conflicting Institutional Logics. Journal of Information Technology, 22, 3, pp. 235-248.

[11] DeLuca, J & Enmark, R. 2000. "E-Health: The Changing Model of Healthcare," Frontiers of Health Services Management, Vol. 17, No. 1, pp. 3-15.

[12] Garg et al. 2005. Effects of computerized clinical DSS on practitioner performance and patient outcomes: A systematic review, JAMA, 293, 10, pp. 1223-1232.

[13] Hair, J., Black, W., Babin, B., Anderson, R., and Tatham, R. 2006. Multivariate Data Analysis, 6th ed. Pearson Prentice Hall, Upper Saddle River, New Jersey.

[14] Harris Interactive, 2002; Boston Consulting Group, US.

[15] Heeks, R. 2006. Health information systems: Failure, success and improvisation. International Journal of Medical Informatics, 75, pp. 125-137.

[16] HIMSS, 2003. "Section II: National Preparedness & Response (NPR)" and "Section IV:National Health Information Infrastructure (NHII)," HIMSS Advocacy Dispatch, www.himss.org.

[17] Hoffman S, Podgurski A. 2009. E-Health Hazards: Provider Liability and Electronic Health Record Systems. Berkeley Technology Law Journal. 24, 4, pp. 1523-1581.

[18] Kolkowska, E., Hedstrom, K., & Karlsson, F. 2009. Information Security Goals in a Swedish Hospital. In G. Dhillon (Ed.), Proceedings of the 8th Annual Security Conference Discourses in Security Assurance and Pri-vacy, pp. 16,111.

[19] Kaplan, S and Garrick, B.J. 1981. On the Quantitative Definition of Risk, Risk Analysis, 1, 1, pp. 11-27.

[20] Karsh et al. 2006. A human factors engineering paradigm for patient safety: designing to support the performance of the healthcare professional, Qual Saf Health Care. 1 pp. i59–i65.

[21] Lohr, H; Sadeghi, A; Vishik, C; Winandy, M. 2009. Trusted privacy domains { challenges for trusted computing in privacy-protecting information sharing. In Information Security Practice and Experience, 5th International Conference, (ISPEC'09), vol 5451 of Lecture Notes in Computer Science, pp. 396-407.

[22] Lorence, D. and Churchill, R. 2005. Incremental adoption of information security in health-care organizations: implications for document management. Information Technology in Biomedicine, IEEE Transactions on, 9, 2, pp.169–173.

[23] Schabetsberger, T; Ammenwerth, E; Andreatta, S; Gratl, G; Haux, R; Lechleitner, G; Schindelwig, K; Stark, C; Vogl, R; Wilhelmy, I; Wozak, F. 2006. From a paper-based transmission of discharge summaries to electronic communication in health care regions. International Journal of Medical Informatics, 75, pp. 209-215.

[24] Seaton B. 2007. Managing e-health Risks and Opportunities. In Proceedings of Annual OACCAC Conference (Ontario Association of Community Care Access Centres), Toronto, Ontario.

[25] Smaltz, D. H., Carpenter, R., & Saltz, J. 2007. Effective IT governance in healthcare organisations: a tale of two organisations. International Journal of Healthcare Technology Management, 8, 2, pp. 2-17.

[26] Trudel, M.; Paré, G. & Laflamme, J. 2012. Health information technology success and the art of being mindful: Preliminary insights from a comparative case study analysis. Health Care Management Review, Vol 37, Issue 1, pp 31–42.

[27] Tsiknakis, M., and Kouroubali, A. 2009. "Organizational Factors Affecting Successful Adoption of Innovative E-health Services: A Case Study Employing the Fitt Framework," International Journal of Medical Informatics, 78, 1, pp. 14.

MEDICAL IMAGES ANALYSIS IN CANCER DIAGNOSTIC

Jelena Vasiljević[1], Ivica Milosavljević[2], Vladimir Krstić[1], Nataša Zivić[3], Lazar Berbakov[1], Luka Lopušina[5], Dhinaharan Nagamalai[4] and Milutin Cerović[5]

[1]Institute Mihajlo Pupin, University of Belgrade, Belgrade, Serbia
[2]Center for Pathology and Forensics of the Military Medical Academy in Belgrade, Belgrade, Serbia
[3]University of Siegen, Germany
[4]Wireill, Australia
[5]The School of Computing, University Union

ABSTRACT

This paper shows results of computer analysis of images in the purpose of finding differences between medical images in order of their classifications in terms of separation malign tissue from normal and benign tissue. The diagnostics of malign tissue is of the crucial importance in medicine. Therefore, ascertainment of the correlation between multifractals parameters and "chaotic" cells could be of the great appliance. This paper shows the application of multifractal analysis for additional help in cancer diagnosis, as well as diminishing. of the subjective factor and error probability.

KEYWORDS

Multifractal, Fractal, Medical Images, Cancer

1. APPLICATION OF MULTIFRACTAL ANALYSIS IN IMAGE PROCESSING

Multifractal analysis of images is based on the definition of measurements with images that are gray levels. Then the multifractal spectrum is calculated. In contrast to many classic approaches, there is no filtering. The spectrum uses local as well as global information for segmentation, noise reduction or edge detection at image points.

Image analysis is a fundamental component of a computer visual problem, with applications in robotics, medical or satellite images ... Segmentation is an important step that provides a description of a basic individual process. Filtering then gives signal gradients where extremes roughly correspond to contours. Then, multi-resolution techniques can be used to "purify" the

Dhinaharan Nagamalai et al. (Eds) : ITCSE, WiMo, ICAIT, ICDIPV, CRYPIS - 2018
pp. 67–77, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80707

results obtained. The main drawback of this approach is loss in precision due to preliminary filtering.

An alternative approach is to observe that a picture is a measure known as a fixed resolution. The irregularities of this measure can then be studied using a multifractal analysis. The general principle is the following: first, the different dimensions and capacitance are defined from the image which is the gray level. Then, the corresponding multifractal spectrum is calculated, providing both local (over) and global (through) information. There are no hypotheses about the regularity of the signal. Multifractal analysis (MF) can be successfully used in image processing. The idea of applying (inverse) MF analysis in extracting characteristic details in the picture is presented in [1].

The importance and the advantage of fractal and multifactal analysis (MFA), in relation to the "classic" signal analysis, lies in the way in which irregularity is considered. The MFA tries to extract information directly from singularity, while in the "classic" mode, most often, NF filtered versions are viewed, possibly with different filtering depths, to detect irregularities and suppress noise. In particular, based on a certain value i, the points of inhomogeneity in the original signal can be separated [1, 2, 3, 4]. By dividing pixel images that satisfy the selected parameter value, or spectrum, by inverse multifractal analysis (IMFA), it is possible to extract from the image of a region that can not otherwise be noticed by any of the known methods. An additional advantage is that such a segmentation does not cause any degradation of the initial image: all the mutual relations of the pixels remain unchanged, so that the details of the image are kept completely. This feature is particularly important in medical diagnostics, so the potential of IMFA in this area is high.

It is shown that a large number of frequently variables of a different nature (electrical signals, modern telecommunication traffic, meteorological and biomedical signals) can be described in a similar way. It is necessary to examine the fractal characteristics for the expression of significant variability. The use of classical statistical methods in such a case (mean value) could cause error estimates. The pronounced singularities indicate the multifunctionality of the process.

2. FRACTAL MORPHOMETRY APPLIED TO TUMORS

Despite the huge increase in our understanding of the molecular carcinoma mechanism, most diagnoses are still determined by visual examination of radiographic images, microscopic and biopsy patterns, direct examination of the tissue, and so on. Usually, these techniques are applied in a quality manner by clinicians who are trained to classify images showing abnormalities such as structural irregularities or high indications for mitosis. A more qualitative and reproducible method, which can serve as an ancillary tool for diagnostic training, is an analysis of images using computer tools. This lies in the potential of fractal analysis as a morphometric measure of irregular structures that are typical of tumor growth.

Pathologists are skilled in examining the boundary surface of the epithelial-connective tissue, which separates the tumor and surrounding healthy tissue. The nature of the tumor edge, whether infiltrative or invasive or poorly expansive, provides information useful not only for prognosis, but also for diagnosis (either benign or malignant tumors). In the study of Landini and Ripini [5], the border area of the epithelial-connective tissue of oral mucosa was examined. Lesions are classified by routine diagnosis into four categories: a) normal; b) medium dysplasia; c) moderate

to severe dysplasia; d) carcinomas. Fractal lesion analysis, which subsequently followed, revealed the following fractal dimensions for the above four categories: 1.07 ± 0.05, 1.08 ± 0.09, 1.16 ± 0.08, and 1.41 ± 0.08, respectively. Although the differences were not large enough to be accepted as an independent tool in diagnostics, they are regardless of consistent measurements of the degree of distortion of the boundary surface. Landini and Ripini then proceeded to describe using a more sophisticated multifunctional analysis method that gives a spectrum of fractal values instead of one value for each image. This method has provided more reliable discrimination of the pathological conditions of the tissue. Lefebre and Benali [6] and Polman et al [7] have shown that fractal methods can also be useful for analyzing digitized mammograms, increasing the hope that the number of incorrect positive mammograms will be reduced in this way. Considering that the increase in irregularities, with associated fractal fracture enlargement, is a common indicator of tumor growth this undoubtedly represents a universal result. Out of everything contained in this chapter Fractals in biomedical systems, the hypotheses set out in this paper are derived.

2.1. Histopathological characteristics of normal mucous membrane, adenoma and adenocarcinoma of the colon

The normal mucous membrane of the mucous membrane of the colon is from the lumen of the intestine to the surface of the laminae epithelialis tunicae mucosae consisting of a layer of cylindrical goblet cells, absortive, endocrine and undifferentiated cells lying on the basal membrane and invaginating the so-called crypts They reach the next layer, laminae muscularis mucosae built of smooth muscle cells, and underneath there is a lamina propria of tuniciae mucosae with blood vessels, loose connective tissue, and individual mesenchymal cells.

The basic histological characteristics that point to the normality, regularity are: uniformity of the cells of the liminal epithelialis tunicae mucosae, polarization of the sails to the basal memebran, small sails, preserved secretion of the mucin by the goblet cells. Figure 1 shows a photo of a sample of the normal colon tissue, he (hematoxilieosin-coloring method), a transverse cross sectional view, and in Fig. 2, photographs of a sample of the normal tissue of the column, he is a longitudinal sectional view.

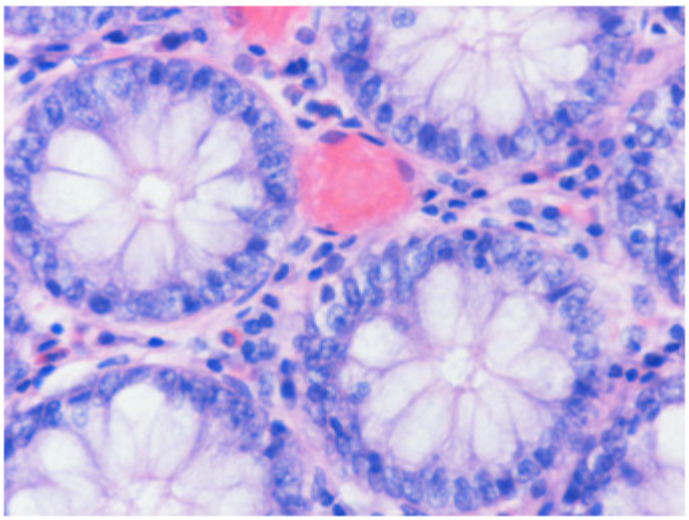

Figure 1: Photo of a sample of normal colon tissue, (he), a transversal section view

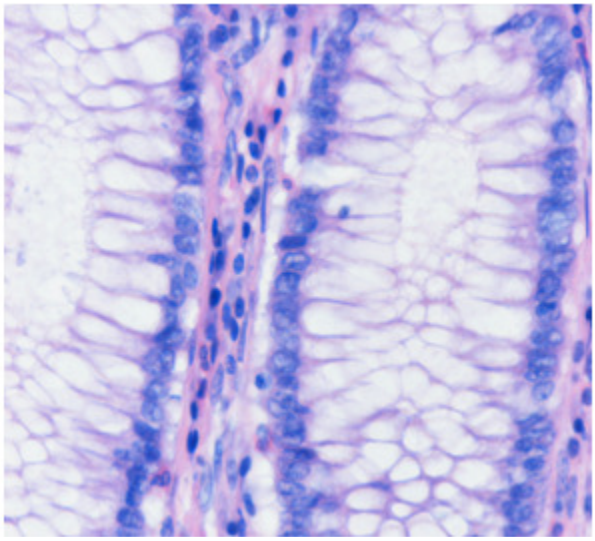

Figure 2: Photo of a sample of normal colon tissue, (he), a longitudinal section view

The most common tumors in the colon are the origin of the epithelium and they can be benign-nature-adenomas and malignant-nature-adenocarcinomas. Adenoma is a benign tumor of the epithelial origin in which proliferation of the epithelium occurs with the formation of glandular structures, which are coated with dysplastic cells. The disposable epithelium loses the uniformity of its cells, the cells are pleomorphic, and the usual architectural appearance is lost, tubular and / or vilosic structures are formed which are coated with cylindrical epithelial cells, with elongated and hyperchromatic nuclei and which may and may not have preserved mucigen activity, while the mitotic activity of the cells is increased. In Figure 3, a photo of the sample of the adenoma column is shown, he is a transversal sectional view, and in Figure 4 a photo of the sample of the adenoma column is shown, and he shows the longitudinal cross section.

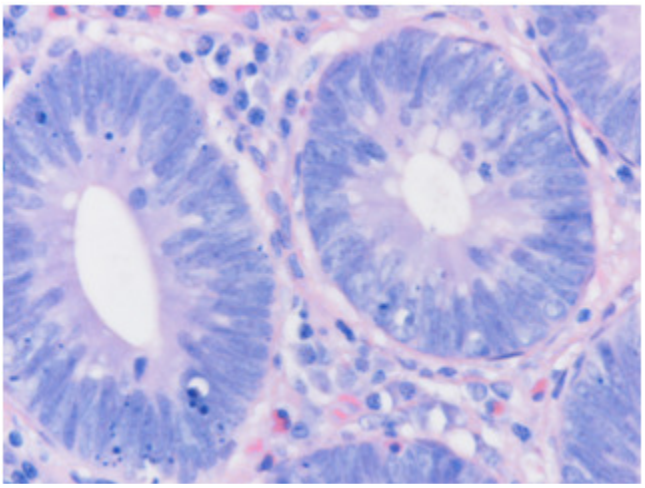

Figure 3: Photo of the column adenoma sample, (he), a transverse cross sectional view

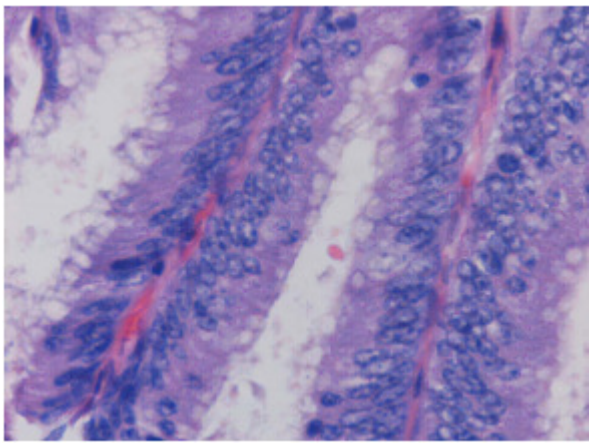

Figure 4: Photo of the column adenoma sample, (he), a longitudinal section view

Adenocarcinoma of the colon is one of the most common tumors in the human population and is one of the major challenges of human medicine, precisely because it can arise from the adenomas and what produces the symptoms relatively early, which allows diagnosis and treatment. Carcinoma tissue is constructed of tubular formations, irregular shape and size, as well as from cribriform formations, which are coated with cubic and cylindrical atypical cells whose nuclei are pleomorphic, hyperchromatic, with and without prominent cats. The surrounding stroma of the tumor is desmoplastic, with multiplied binders and with inflammatory lymphocyte infiltrate, plasmocyte, and granulocyte. The properties of adenocarcinoma are: local invasive growth by penetrating the basal membrane, infiltrating the column wall and possibly spreading into the surrounding structures, as well as the ability to metastasize to regional lymph nodes, and metastasis to distant organs: in the liver, lungs, bones, and other organs [CECI89]. Figure 5 shows a photograph of a sample of colon carcinoma, a he-transversal sectional view, and Figure 6 shows a photo of a sample of colon carcinoma, he- an indication of the longitudinal cross-section.

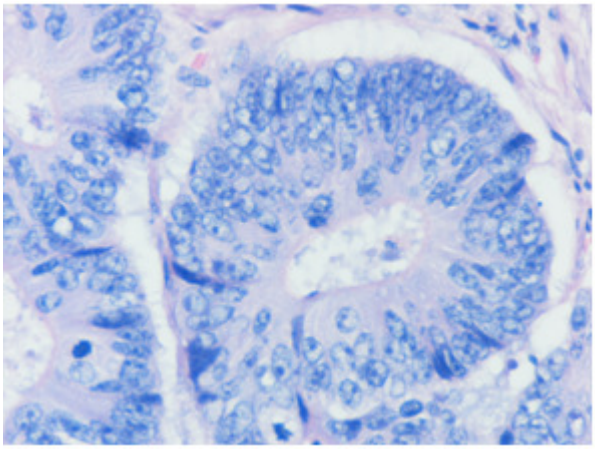

Figure 5: Photo of the column carcinoma sample, (he), a transverse cross-sectional view

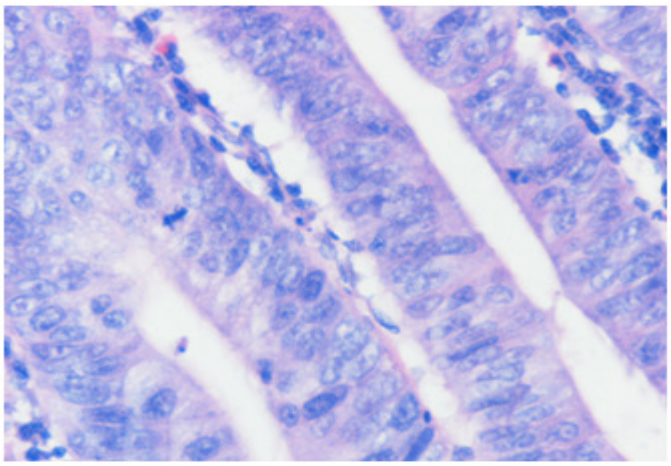

Figure 6: Photo of the column carcinoma sample, (he), a longitudinal section view

2.2. Methodological part of the research

Research goals

The intention behind this research is to determine the existence of differences in the parameters of the multifractal analysis of digital medical images between the following three tissue groups:

1. Normal mucosal tissue of the colon
2. Thick bowel mucosal tissue with diagnosis of malignant tumor origin of the epithelium – carcinomas
3. Thick bowel mucosal tissue diagnosed as benign tumors of the origin of the epithelium – adenomas

In order to obtain the required research results, it is necessary to perform the following steps:

1. Determine the parameters of the multifaceted analysis for the previously stated groups of tissue pictures
2. Determine the existence of statistically significant differences between the parameters corresponding to the previously mentioned tissue image groups.

2.3. Method

It is a non-experimental correlation study on the sample. In relation to the goal, the research is parametric.

2.4. Variables of research

Independent variable

1. Type of photographed tissue. Variables are operatively defined with the following categories:

a) tissue without pathological changes

b) tissue with cancer

c) adenoma tissue

Dependent variables

In the FracLac program

1. $D_{\max}$ - maximum
2. $\overline{Q}$ -Q which corresponds to the maximum
3. $\underline{\alpha}$ - which corresponds to the minimum
4. $f(\alpha)_{\min}$ - minimum
5. $\overline{\alpha}$ - which corresponds to the maximum
6. $f(\alpha)_{\max}$ - maximum

In FracLab

1. α_{sr} - mean value
2. $f(\alpha)_{sr}$ - mean value
3. $\overline{\alpha}$ - which corresponds to the maximum
4. α_{stdev} - standard deviation
5. $f(\alpha)_{stdev}$ - standard deviation

2.5. Instruments

The following programs were used for multifractal analysis of the obtained digital medical images and obtaining the parameters of multifactal analysis:

1. "FracLac" program for multifractal image analysis [8].
2. "FracLab" program for multifractal image analysis [9].
3. Program for statistical analysis of data SPSS (Statistical Package for Social Sciences), standard for statistical analysis of clinical research results.

2.6. FracLac

The "FracLac" program is created in the Java programming language and represents one of the plugins in the "ImageJ" digital image analysis program. "ImageJ" is freely available image analysis software written in the Java programming language authored by Wayne Rasband of the National Institute of Health of the United States of America from Bethesda, Maryland, [8].

The author of "FracLac" is Audrey Karperian from Charles Sturt University in Australia. She was contacted with this research. At her request, they were sent one sample sample from each of the three groups to test and improve the program for the specific problem of this research. With great help from her professional team, this program is adapted to handle a large number of images at once, relatively quickly. This is especially suited for use by medical personnel, where it would not be complicated training, nor would it take up a lot of their time. The images necessary for processing can be collected and then all at once processed.

2.7. The program "FracLab"

The FracLab program was developed by a team of experts at the INRIA and IRCCyN Institutes in France, led by Jacques Levy Vehell.

The principle of image segmentation using multifactal analysis is as follows: the points lying in the picture can be classified according to their Holder exponent. Let's look at the example of the points that lie on the contours. These points often correspond to the discontinuities of the gray level map or from its output. They therefore generally have a "low" Holder regularity. However, the exact value of the exponent will depend on the characteristics of the image. Additionally, the boundary edge feature is not purely local, and therefore a global criterion is required to decide whether a point is assigned, a point that belongs to the edge. Indeed, the points lying on the textures of the region also have a generally low regularity, and it is necessary to find a way to make them different from contours. Here, the other component of the multifractal analysis comes to the fore: since the edges are by definition the sets of points of the dimension one, we declare that the point lies on the contour if there is an exponent such that the associated value of the multifractal spectrum is one. In addition to the geometric characterization of the edges of the edges, it is also possible to make statistical points: the points of the edges can be defined by their probability of being affected when the pixel is randomly selected in the image at a given resolution. The relationship between the geometric and statistical representation of the edges of the edge provides multifractal formalism. Instead of edge detection, a much more complicated structure can be extracted using the same principle: starting again from Holder's exponents, points can be kept where the spectrum has a certain value. For example, by selecting a value of about 1.5, it is generally possible to extract very irregular contours. The value close to 2 corresponds either with smooth regions or textures.

The general procedure is as follows: it begins by calculating the Holder exponent at each point. This gives a picture of Holder's exponents.

The second step is calculating the multifunctional spectrum. In this paper the Hausdorf spectrum is calculated from the three spectra offered. The Hausdorf spectrum gives geometric information that relates to the dimension of the set of points in the image with the given exponent. This spectrum is a function in which apscis represents all Holder's exponents appearing in the image, and the ordinate is the dimension of the set of pixels with the given exponent.

The second spectrum is a large deviation spectrum that gives statistical information related to the probability of finding a point with a given exponent in the image (or more precisely, as this probability acts in the change of resolution).

The third spectrum is the so-called. The Leandre's spectrum, which represents only the concave approximation of the spectrum of large deviations, and its main contribution is to give much more robust calculations, although at the cost of losing information.

2.8. Hypotheses

General hypotheses

The parameters of multifractal analysis will significantly differ from all three groups: normal tissue, carcinomas and adenomas.

1. In particular, in the case of multifractal analysis of digital images of three groups observed, FracLac will distinguish the following obtained parameters for all three groups observed:

- $D_{\max}$ - maximum

- $\overline{Q}$ - Q that corresponds to the maximum

- $\underline{\alpha}$ - which corresponds to the minimum

- $f(\alpha)_{\min}$ - minimum

- $\overline{\alpha}$ - which corresponds to the maximum

- $f(\alpha)_{\max}$ - maximum

2. In the case of multifractal analysis of digital images of the three groups observed, FracLab will distinguish the following obtained parameters for all three groups observed:

- α_{sr} - Middle value

- $f(\alpha)_{sr}$ -Middle value

- $\overline{\alpha}$ - which corresponds to the maximum

- α_{stdev} - standard deviation

- $f(\alpha)_{stdev}$ - standard deviation

2.9. Sample

The sample consisted of 150 preparations obtained from biopsy from the gastrointestinal tract, more precisely from the colon, Adenocarcinoma tubulare coli. Of the 150 preparations, 50 were previously diagnosed as normal colon mucosal tissue, 50 were diagnosed as colon mucosal tissue with malignant epithelial tumors - carcinomas, and 50 diagnosed as colon mucosal tissue with benign epithelial tumors - adenomas.

In the Center for Pathology and Forensics of the Military Medical Academy in Belgrade, the preparations were prepared for analysis under a microscope with a magnification of 40x and photographed on a coolscope device (by the author with the help of Dr. Ivana Tufegčić) in digital form (Figures 1, 2 and 3). Coolscope is a kind of hybrid microscope in the body of the computer, the manufacturer is a Japanese firm, Nikon. Five different photographs were taken from each preparation, in order to obtain the most valid results of statistical analysis. In this way, a total of 750 digital images were obtained, 250 of each of the three groups.

3. CONCLUSIONS

In this study, multifractal analyzes were performed using two programs FracLac and FracLab, three groups of tissue pictures: normal tissue, carcinoma, and adenomas. Then, statistical processing of the obtained results was made using the SPSS statistical treatment program, which is usually used in clinical trials. In this way, the answers to the hypotheses posed in this paper are obtained. The general conclusion is that the basis of the general hypothesis proved to be correct, that the parameters of multifractal analysis differ significantly for all three groups of tissue tissues observed, and therefore the zero hypothesis about the non-separation of these groups was denied. This applies when applying both programs for multifractal analysis of FracLac and FracLab images.

Since the general hypothesis of this paper is confirmed, it can be concluded that this research has obtained positive results.. In the case of the FracLac program, the reliability of the classification of all three tissue groups analyzed based on the obtained multifractal parameters is 65.3%, which is more than the 60.7% obtained in the case of FracLab. In the case of the FracLac program, 80.0%, 73.0% and 85.0% were obtained, successfully classified cases for the following group relationships (respectively): normal tissue and carcinomas, carcinomas and adenomas, normal tissue and adenomas.

In the case of the FracLab program for the same relationships, tissue image groups, respectively, received the following reliability of 64.0% (which is considerably worse in relation to the same case with the FracLac program), 74.0% (which is a little better in relation to the same case for the FracLac program) and 80.0% (which is worse in relation to the same case with FracLac), we can conclude that the resulting classifications are very effective and this is generally more in the case of FracLac.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education and Science, Republic of Serbia, Science and Technological Development grant TR32037.

REFERENCES

[1] J. L. Vehel, (1996) "Introduction to the multifractal analysis of images", http://www-rocq.inria.fr/ fractales

[2] B. Reljin, I. Pavlović,I. Reljin, I. Rakočević, (1999) "Multifraktalna analiza medicinskih slika", Zbornik VII konf. TELFOR-99, page.469-472

[3] P. Mannersalo, I. Norros, (1997) "Multifractal analysis: A potential tool for teletraffic characterization?", COST257TD(97)32, pp.1-17

[4] P. Mannersalo, I. Norros, (1997) "Multifractal analysis of real ATM traffic: A first book", COST257TD(97)19, pp. 1-8

[5] Landini G., Misson G. P., Murray P. I., (1993) "Fractal analysis of the normal human retinal fluorescein angiogram". Curr. Eye Res, 12: 23-27,.[Medline]

[6] Lefebvre F., Benali H., (1996) "A fractal approach to the segmentation of microcalcifications in digital mammograms". Med. Phys, 22: 381-390, 1995.[Medline]eds. Fractal Horizons, : 251-262, St. Martin's Press New York.

[7] Pohlman S., Powell K., Obuchowski N. A., Chilcote W. A., Grundfest-Broniatowski S., (1996) "Quantitative classification of breast tumors in digitized mammograms". Med. Phys, 23: 1337-1345,.[Medline]

[8] Wayne Rasband, ImageJ, http://rsb.info.nih.gov/ij/

[9] J. L. Vehel, FracLab, http://www.irccyn.ec-nantes.fr/hebergement/FracLab/

AUTHOR INDEX