LIVESGEO: A Location-based Multimedia Knowledge Sharing System

by

Yongik Chung

Honours B.Sc., The University of Western Ontario, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies
(Computer Science)

THE UNIVERITY OF BRITISH COLUMBIA
(Vancouver)

August 2011

Abstract

Learning is vitally important because we live and work in a rapidly changing world. As learning is acquisition of knowledge or skills through utilizing given information, it is critical that one needs to be literate in order to learn effectively. However, conventional learning materials are mostly targeted for literate audiences and delivered by the means of written materials in classical educational settings making it harder for illiterate audiences to engage in learning. This thesis introduces a novel location-based multimedia knowledge sharing system, named LIVESGEO, which takes advantages of advanced capabilities of smart-phones and mobile data networks to allow peer driven learning. The system allows illiterate audiences to easily create, share, search and rate knowledge contents on the spot. The system proposes a hybrid of client-server architecture and peer-to-peer architecture to efficiently circumvent the critical limitations and leverage the advantages of mobile devices. We also propose a novel mathematical framework to characterize relative popularity of a rated content based on its reputation and quality and justify the performance with empirical evaluations against popular rating frameworks. We also explore the applicability and usefulness of geo-tagged contents accumulated by LIVESGEO. To demonstrate this, we abstract the system into a spatial content cluster framework, which acts as an abstract layer between reverse-geocoded contents and applications to provide information about the contents that are clustered based on their geo-location, to allow unique applications to be built upon the system. In addition, we provide examples of such applications which utilize the spatial content cluster framework. In order to justify the use of reverse-geocoding to cluster contents geographically, we provide empirical analyses of reverse-geocoding by making novel use of two popular mathematical models in the field of cartography, namely the Haversine and Vincenty formulae.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank the faculty and my fellow students at the NIC lab, who have encouraged and inspired me during the research process. I owe particular thanks to my supervisor and mentor Dr. Son. T. Vuong, for his helpful supervision and support. I would also like to thank George Tsiknis for his guidance and my buddies - Andrew Tjia, Ricky Cheng, Shahed Alam, Stanley Chiu, and Jonatan Schroeder. Special thanks are owed to my soul mate Jaeyoung Kim and my family, who always love and believe in me.

Yongik Chung
July 2011

1 Introduction

This section summarizes the question of how to improve open learning, which LIVESGEO is seeking to answer by leveraging mobile and data network technologies, and provides general ideas on why such a question is worthwhile for research.

1.1Motivation

Learning is vitally important because we live and work in a rapidly changing world. New ideas, policies and technologies are expeditiously introduced to amend emerging problems and improve the quality of life; and we are under demands to comprehend and learn from them to grow with the rest of the world. The United Nations Educational, Scientific and Cultural Organization (UNESCO) defines literacy as the "ability to identify, understand, interpret, create, communicate, compute and use printed and written materials associated with varying context [1]." It is clear from the definition that literacy is a key factor for a person to develop one's knowledge and skills. As learning is acquisition of knowledge or skills through utilizing given information, it is critical that one needs to be literate in order to learn effectively. However, conventional learning materials are mostly targeted for literate audiences and delivered by a means of written materials in classical educational settings making it harder for illiterate audiences to engage in learning.

To provide alternative learning contents [1] specifically tailored for illiterate audiences, Learning through Interactive Voice Educational System (LIVES) [2,3] was developed at the University of British Columbia. LIVES offers asynchronous two-way communication learning where the voice-based learning contents are unidirectionally delivered to the students. However, strictly adhering to voice-based learning limits the capability of LIVES as an effective educational system. To amend this limitation, LIVESMOBILE [4] was recently developed to leverage the MMS capability of mobile phones to enhance the learning experience of LIVES. LIVESMOBILE can send short video clips to the students via a MMS gateway. However, the size of a MMS video-based learning content is heavily limited by the inherent size restriction on MMS. For example, T-Mobile [5] only allows 1MB MMS to be sent over its networks, which hinders the whole idea of using MMS video as a learning content.

[1] We will be using the term *content* as a countable noun throughout the thesis to represent a single learning content module.

This thesis introduces LIVESGEO which amends the aforementioned limitations of both LIVES and LIVESMOBILE and further improves the learning experience by integrating geo-location metadata to learning contents and allowing the students to become a content creator and create multimedia learning contents to be shared over 3G and emerging 4G data networks.

1.2Thesis Contributions

Location-based knowledge sharing in general is only beginning to receive scholarly examination. There have been studies on analyzing patterns of the knowledge sharing systems usage but there has not been any study on integrating location information onto the user created multimedia content and tailoring the system to work effectively on mobile devices. One of the main purposes of this thesis is to layout preliminary design and implementation of a location-based multimedia knowledge sharing system tailored for mobile devices to facilitate future research. Another major purpose is to explore the applicability of the geographical information of contents. To show its applicability and usefulness we abstract the learning system into a framework allowing interesting applications to be built upon the system that utilize the geographical information adhered to the contents. The following highlights the main contributions of this thesis:

1. Introduces a novel location-based multimedia knowledge sharing system that takes the advantages of advanced capabilities of smart-phones and mobile data networks to allow peer driven learning. The system provides an intuitive user interface to allow the users to easily create, share, search and rate geo-tagged multimedia knowledge contents on the spot.
2. Provides high level design specifications and implementation decisions to enable the system to work most effectively as a distributed mobile application. The system proposes a hybrid of client-server and peer-to-peer architectures to efficiently circumvent the critical limitations and leverage the advantages of mobile devices.
3. Provides our own mathematical framework to characterize the true popularity of a content relative to other contents in the system using their raw user rating scores. We evaluate its performance through empirical comparisons with popular rating frameworks.
4. Explores the applicability and usefulness of the geo-tagged contents accumulated by LIVESGEO. To demonstrate this, we abstract the system into a spatial content cluster framework, which acts as an abstract layer between reverse-geocoded contents and applications to provide information about the contents that are clustered based on their geo-

location, to allow unique applications to be built upon the system. We also provide examples of such applications that utilize the framework.

5. In order to justify the use of reverse-geocoding to cluster contents geographically, we provide empirical analyses of reverse-gecoding by making a novel use of two popular mathematical models in the field of cartography, namely the Haversine and Vincenty formulae.

1.3Thesis Organization

The thesis, excluding this introductory chapter and the concluding chapter, is broken down into six main chapters. The second chapter provides the importance of open learning and its socioeconomic impacts to emphasize the valuable motivations that have driven this work. The third chapter provides a brief introduction to development on mobile platform, mobile data network, and data mining technique to allow one to follow this thesis easier. The fourth chapter provides related works concerning different styles of knowledge sharing systems and their advantages and limitations. The fifth chapter provides a detail description of design and implementation of LIVESGEO. The sixth chapter provides a mathematical framework to characterize the popularity of rated contents and empirical experiments to evaluate its performance. The seventh chapter explores the applicability of geo-tagged contents and documents the spatial content cluster framework with empirical experiments to evaluate its performance and to justify the methods that have been used to design the framework.

2 Importance of Open Learning

For an individual, literacy is essential to function in today's world because it directly influences one's economical and social well-being. In this section: first, we analyze how literacy is correlated with the important factors of life such as career, quality of life, health, and self-perception and second, we describe the rationale behind the use of technologies that LIVESGEO is built upon.

2.1Illiteracy and its Socioeconomic Impacts

According to National Literacy Trust [6], 92% of the British public consider literacy as a vital skill to find a job and be successful. Another study by the same research group [7] has found that men and women with poor literacy are least likely to be in full-time employment at the age of thirty, and 63% of men and 75% of women with very low literacy skills have never received a promotion. These facts emphasize that poor literacy can become a serious barrier to progress in their career even after successful employment. In a competitive global economy, highly skilled workers are an asset to economic growth at a national and global level. The advancement of sophisticated communications and computing technologies have been leading the occupational composition in the world to shift toward the occupations requiring a higher level of education, in turn, naturally leading the countries with a higher literacy population to have higher GDP-per capita [8]. The studies in Canada have shown that only a 1% increase in an average literacy rate can yield a 1.5% or $18 billion increase in the GDP in Canada [9] and also that literacy skills have significant influences on the probability of being employed and average income [10].

Literacy has a direct correlation to health issues. New England Journal of Medicine [11] reported that the people with low literacy are more likely to report having poor health, and are more likely to have diabetes and heart failure than those with adequate literacy. Moreover, the doctors have experienced that illiterate patients are irregular in follow up and difficult to ascertain if they have taken the medication as prescribed. This may due to the fact that health care systems assume that patients are literate and thus will be able to comprehend their medical conditions and explain their concerns.

The study in [7] highlights the evidences that literacy has a direct correlation to one's quality of life and poor literacy disadvantages adult lives. The study points out a high divorce rate among the illiterate population compared to the literate population. It also profiles other interesting socioeconomic impacts of illiteracy: A person with poor literacy is less likely to vote, more likely to live in a overcrowded house, less likely to own their home, less likely to participate in their community, less likely to trust people in their community, more likely to perceive their community to be an unsafe place, more likely to smoke and more likely to have poorer mental health.

It is safe to say that literacy is a minimum prerequisite for succeeding in today's globalized and knowledge demanding world, with the supporting facts of the aforementioned studies. However, many people in the world, even in highly developed countries like the United States, Canada and Europe, are still illiterate. UNESCO Institute for Statistics [12] reported in 2004 that 800 million people are illiterate, which is nearly 1 in 6 people in the world, and these populations are concentrated in developing regions of the world. Many studies have pointed out the strong link between illiteracy and poverty as the main cause of illiteracy. There is a significant negative correlation between poverty and adult literacy rates [12]. The regions in the world where poverty rates are higher tend to show lower literacy rates. This is due to low-income households having less access to higher-quality education and non-formal education programs. High-income households also tend to reside in more literate environments, which are naturally more demanding of literacy skills. Moreover, high-income households who are literate are more likely to be able to support their children for their educations increasing the chance of their children becoming highly-literate. Therefore, UNESCO has been proposing that open learning environment must be developed and accessible for educations and trainings to any individual, especially to those who are in need, to weaken the link between literacy and poverty.

2.2 Approach to Open Learning: LIVES, LIVESMOBILE, and LIVESGEO

LIVES [2,3] is an open learning software system developed as a solution to decrease the illiteracy rates around the world by offering an open learning environment to illiterate people in developing countries, leveraging the mobile and cost effective VOIP technologies. LIVES takes the advantage of a recent phenomenon in developing countries, which is the increase of mobile phone penetration rates, and offers voice-based lessons specifically designed for illiterate students on their mobile phone. LIVES offers a bidirectional communication channel where the students can listen to the voice-based learning material and respond to the questions. LIVES also provides asynchronous learning where the user can learn at one's preferred time. However, strictly adhering to voice-based learning content limits the capability of LIVES as an effective educational system. Learning can be better facilitated by use of images, diagrams and animations. To amend the voice only limitation of LIVES, LIVESMOBILE [4] was recently developed to leverage the multimedia capability of mobile phones to enhance the learning experience of LIVES. LIVESMOBILE integrates Multimedia Message Service (MMS) into the LIVES system to deliver short video-based learning materials to the students. However, the inherent size restriction on MMS enforced by service providers hinders the whole idea of using video as an effective learning content. Furthermore, LIVES and LIVESMOBILE limit themselves to the expert-driven learning model, where learning contents are created by the experts only and delivered unidirectionally to the students. Unlike LIVESGEO's peer driven model, the expert-driven model does not allow knowledge to be created and shared by the users themselves, which is proven to be effective by popular knowledge sharing systems such as Wikipedia.

Both LIVES and LIVESMOBILE are designed based on the assumption that developing countries have only limited access to high-end technologies such as smart-phones and mobile data network. However, recent statistics have reported increasing rates of global mobile data traffic usage and smart-phone penetration, which are even higher in developing countries. Cisco [13] forecasted that global mobile data traffic will double every year through 2014 and increase 39 times between 2009 and 2014. Global mobile data traffic increased 160 percent from 2008 to 2009. Cisco also reported that the key driver of mobile data traffic is the smart-phones capable of advanced multimedia capabilities. The global smart-phone penetration level is currently 9% and expected to grow to 17% by 2014 [14]. Although developing countries' current penetration rates are relatively low compared to developed countries, its estimated growth rate is much higher, which is 241% by 2014 compared to 182% of developed countries. Another global mobile statistics [15] reported that half a billion people accessed the mobile Internet worldwide in 2009 and the usage is expected to double within five years and almost one in five global mobile subscribers have access to the high-speed mobile Internet. As LIVESGEO takes a different approach from LIVES and LIVESMOBILE by leveraging the advanced features of today's smart-phones such as multimedia and GPS capabilities, and high speed mobile data network, LIVESGEO can encompass much wider audiences in both developed and developing countries. Moreover, LIVESGEO will become more attractive as the penetration level of smart-phone and mobile data network is forecasted to steadily increase in the future.

3 Background Information

This chapter provides a brief overview of mobile development and Android platform, which is the mobile open source platform LIVESGEO is built upon, and the fundamentals of mobile data network, and data mining techniques. This background information will provide the readers with some measure of background information relevant to a location-based multimedia mobile knowledge sharing system.

3.1Brief History of Mobile Development and Android Platform

Previously, mobile development was a closed environment led only by the developers of mobile device manufacturers built on proprietary operating systems requiring proprietary development tools [16]. Mobile phones prioritized manufacturers' native applications and third party applications had only limited access to devices' features, and that is if the third party application development were allowed to begin with. Often, developers needed to code in low-level requiring the understanding of the hardware specifics of devices. Nokia first pioneered platforms, namely Symbian, for developers to write an application once and run on a wide variety of the Nokia devices. However, it was not successful in attracting and increasing the population of mobile developers due to the high learning curve of development requiring the use of proprietary Symbian C++. In more recent years, Java introduced Java Mobile Edition (Java ME), which runs on any mobile device that has Java Virtual Machine installed. However, it failed to attract interest from hardware manufacturers to have Java ME installed on their devices. Thus, hardware providers only allowed very limited access to the native hardware features for Java ME to take advantage of. These barriers limited Java ME from becoming a mature and popular mobile development platform.

Android [17,18] is an open source mobile operating system and application platform by Google that is made up of Linux kernel based operating system, middleware, and software development kit. According to Wireless Federation report [19], the global market share of Android is currently 35% making it the world's most-used smart-phone platform and it also has the largest community of mobile developers. Android also has the second largest mobile application ecosystem, next to Apple's iOS, which makes the downloading and updating of third party applications easier through Android Marketplace. It supports basic GSM, CDMA or LTE telephony to more advanced features such as video and audio support, GPS, camera, and Wi-Fi. Being an open source platform, Android offers developers the ability to build OS-level native applications, which means that the developers have access from application level API to operating system native system calls. Simply, developers have full access to the API that the platform itself is built with. A developer can even modify the OS itself. The Android platform runs each application in its own separate application sandbox making each running instance of an application tightly secured. Although, Android is similar to Java at its core, it distinguishes itself by having a unique virtual machine called Dalvik which is optimized for efficient mobile device performance allowing the applications to leave as minimal memory footprint as possible. Android is known to be superior to other mobile platforms in many features distinct to mobile platforms such as portability, reliability and openness [20].

3.2Basics of Mobile Data Network

There are numerous mobile data network technologies and standards developed by different groups over the world. However, for the sake of simplicity, this section will focus on describing the most popular GSM network. Global System for Mobile Communication (GSM) in an international wireless communication standard and it is the most popular and ubiquitous standard for cellular networks. GSM was originally designed for voice telephony and later evolved into Universal Mobile Telecommunication System (UMTS), commonly known as 3G, to support packet data transport and only recently into Long Term Evolution (LTE) commonly known as 4G, to support high-speed data transfer. With current 3G and 4G technologies, we can surf the Internet and transfer traffic demanding data, such as a high definition video, at the speed of 28Mbps download and 11.5Mbps upload for 3G and 326Mbps download and 86Mbps upload for 4G [21]. However, their empirical performances vary on numerous factors such as a service provider's network configurations and software, and hardware equipment [22]. UMTS, the 3G network, is built upon the existing technologies of GSM. It is based on a packet based technology, much like the Internet, which means that data is queued up in one end and sent only when it becomes a discrete packet through the network. This avoids network resource being tied up during the whole communication process between the end points. LTE is the emerging standard for the latest high-speed mobile network technology. The main goal of this technology is to improve the speed and coverage of the existing 3G networks and there have been various pragmatic and research proposals to achieve this goal. Generally, LTE is designed to optimize itself by taking advantages of the current advanced topology networks and heterogeneous communication networks [23]. These technologies evolved simple voice telephony networks into today's multimedia mobile data networks, where we can provide and consume richer quality information and traffic demanding multimedia data under wider communication coverage.

3.3Overview of Data Mining

Data mining is all about uncovering patterns in data. In essence, data mining is nothing new. What is new is the staggering amount of increasing data and the opportunities to find interesting patterns, requiring more sophisticated and efficient techniques and algorithms. The patterns from a large dataset can be used to solve many real-world problems, such as marketing, profiling, and scientific predictions. What makes pattern finding hard and also interesting in real-life is that datasets are usually incomplete, the values of some features, or attributes, may be missing or unknown and there usually is noise or error in the data. Thus, the challenge is to uncover and generalize the pattern in the data. In essence, data mining technologies offer two important capabilities: the prediction of trends and behaviors and the discovery of unknown or hard to find patterns. LIVESGEO attempts to discover geographical patterns from geo-tagged multimedia contents using a particular data mining technique called clustering, designed to cluster instances into groups according to similarities in their attributes or features.

Clustering techniques are applied when there is no class to be predicted but rather when the instances are to be divided into groups. The clusters "presumably reflect some mechanism at work in the domain from which instances are drawn, a mechanism that causes some instance to bear a stronger resemblance to each other than they do to the remaining instance [24]." For example, clustering can identify the geographical relationships in a location-based dataset, where each instance has geographical metadata, which might not be logically derived or take excessively long time through casual observation. Clustering is abstractly represented by Venn Diagrams where each instance falls into one specific cluster group or overlapping regions of two or more cluster groups or into all available groups with different probabilities. A clustering algorithm trains itself to cluster an incoming instance strictly from the relationships that exist in the previous set of instances before the incoming instance. Thus, the prediction accuracy and performance largely vary depending on the quality of the initial instances available to train the algorithm.

4 Related Work

In the following section, we present popular knowledge sharing systems that share the same goal with our system. We start from the most simple, but also the most celebrated, text based knowledge sharing system, Wikipedia. Then we delve into a multimedia web based knowledge sharing system called digitalGreen, specially designed for developing countries. Finally, we provide an overview of a successful knowledge sharing system called UpSide Learning that utilizes the power of mobile devices.

4.1Text-centric Knowledge Sharing System: Wikipedia

Wikipedia is a popular, free, collaborative, multilingual open learning encyclopedia ranked as the seventh most visited website in the world [25]. It is the largest and most popular knowledge sharing website on the Internet [26,27] composed of 18 million articles written by peer editors and 356 million active readers. According to an expert-led investigation carried out by *Nature* [28], the quality and reliability of the current knowledge contents in Wikipedia is even comparable to the highly reputed Encyclopedia Britannica; their evaluation revealed that the difference in accuracy was not significant.

Similar to the concept of LIVESGEO, Wikipedia departs from the expert-driven style of knowledge sharing and adapts to the peer-driven style, where anyone with access to the service can create new articles and edit existing articles. The peer-driven style allows rapid information creation, update and sharing. It is purely web based where the users can manipulate the knowledge contents on a web browser, where the contents are mostly text-based with some inclusion of illustrations. The popularity of Wikipedia is a proof that open learning can be profoundly successful and attracts interest from individuals to governments despite the fact that the knowledge content creator or editor does not monetize the contents.

However, the strength of peer-driven content sharing is also the major weakness of Wikipedia. There are concerns that the contents from anonymous contributors are prone to misinformation and information vandalism [26]. To address this concern, researchers have suggested that the information vandalism in knowledge sharing systems can be reduced and even prevented by enforcing community introspection [29,30]. In other words, they suggest that the best way to circumvent information vandalism and misinformation is to make the users to pay attention and care about their own contribution. LIVESGEO takes the suggestion from the aforementioned studies and tries to create community surveillance to respond rapidly to information vandalism by integrating a rating system that rewards more informative and high quality knowledge contents. This would also give a content creator a sense of achievement when one's content becomes more popular as it gets higher ratings. Other limitations of Wikipedia are its web-based interface for content management, which is designed primarily to be edited and viewed on conventional monitors and XML like syntax enforced by the system to be followed when creating contents. These make Wikipedia less attractive as a knowledge sharing system on mobile devices.

4.2Multimedia-centric Knowledge Sharing System: DigitalGreen

DigitalGreen [31,32,33] is a multimedia knowledge sharing system designed specifically for developing countries. Its current aim is to deliver agriculture related videos to small farmers in India to enable the farmers to progressively become better farmers and reduce costly expert field support. The farmers perceive that agricultural knowledge is often protected by chemical and seed manufacturers who monetize the knowledge contents. Moreover, government expert field officers are usually unable to visit the farms for economical and political reasons. Thus, the farmers tend to rely on intuition and the hearsay of local townspeople which may sometimes jeopardize their living by making uninformed decisions.

DigitalGreen follows two main principles. First is cost realism: the system should be designed so that the cost of setting up and using the system is minimized. Second is participatory learning: the learning content should be created to maximize the interaction with the students. The video knowledge contents are created by local instructors with digital cameras. Video is selected as it is the best multimedia medium to capture the actual agricultural scenes in detail. One important factor of the video learning content of DigitalGreen is that it includes the local farmers as actors. In other words, local social networks are tapped to connect the farmers with the DigitalGreen system. The rationale behind this approach is that the excitement of appearing on TV would motivate the farmers to get more involved in the learning process. The video contents are then stored in and managed via the main repository through its special content management system called Connect Online Connect Offline (COCO) [34]. The principal means of content distribution is to distribute the video learning contents as DVD to villages. The contents are viewed at dedicated kiosks setup with DVD player and TV to train a registered group of farmers. COCO, which is a web service made with Google Web Toolkit, Java/JavaScript and SQLite database, is the software that is the heart of DigitalGreen. Its main focus is to make information and video content gathering less error prone, fast, and easy under low and limited bandwidth. Its highlighted feature is its offline-capable data input framework. Contents can be uploaded to the main repository using web browsers even with sporadic internet access.

DigitalGreen has a critical limitation as an open learning system due to the fact that the system is purely expert-driven and asynchronous. In other words, the contents are created only by the experts and delivered only at a specific time to the dedicated kiosks. LIVESGEO takes an opposite approach of allowing the peers to create and share their own learning contents and access them asynchronously at their convenience.

4.3Mobile-centric Knowledge Sharing System: UpSide Learning

UpSide Learning [35] is a learning system that blends a traditional learning management system with mobile learning. The system is targeted for enterprise use, mainly to train employees on the go. It helps organizations to deliver different mobile learning contents through various channels; it offers mobile courses, mobile videos, podcasts, mobile books, interactive quiz applications through built in application market place and application stores of iOS, Android and BlackBerry.

The architecture of the system is composed of a proprietary back-end learning content management system and front-end mobile applications. The learning content management system delivers short video lessons or mobile books to the user's mobile device. The system also provides mobile games which provide short quiz games related to the learning materials. The most unique feature of UpSide Learning from other knowledge sharing systems is its focus on the use of mobile games. The system is designed to maximize the learning process by analyzing when and how to use digital games most effectively to enhance learning [36].

Award winning UpSide Learning is a case in point of how smart-phones and the multimedia contents delivered via a high-speed mobile data network can build a successful and effective learning system. LIVESGEO shares a lot of similarities with UpSide Learning and takes one step forward to associate geographical meta-data with learning contents to provide a unique experience and encourage peer driven open learning.

5 LIVESGEO Design and Implementation

This chapter defines an abstract framework for a location-based multimedia mobile knowledge sharing system and its preliminary implementation to facilitate future research. We provide a general overview of the system and explain important design and implementation decisions.

5.1LIVES and LIVESMOBILE

As LIVESGEO is motivated by the work of LIVES and LIVESMOBILE, it is worthwhile to understand the underlying architectures of both systems and how LIVESGEO evolved from its predecessors.

5.1.1 LIVES Design and Implementation

LIVES is a cost effective educational software system which leverages existing conventional telephone networks to disseminate auditory learning components to illiterate audiences without having to create and invest on new communication channels and equipment. As LIVESGEO allows creation of audio contents on a mobile device, it could be used to create audio based learning contents through its mobile client and deliver the contents via LIVES. LIVES is broken down into two sub-systems:

1. Learning Management System (LMS): The main purpose of the system is to seamlessly disseminate learning materials over communication networks to the students. It deals with how and when the learning materials should be sent to the students.
2. Learning Content Management System (LCMS): The main purpose of the system is to provide an interface for the content distributors to design, organize, and upload new learning contents. It also provides an interface to manage a student's profile and progress.

Figure 1 depicts the overall architecture of LIVES. The main components are:

1. Communication Server: This server is a gateway, built upon the open-source private branch exchange telephony software Asterisk [37], between the LIVES system and the VoIP communication channel. It is responsible for establishing and maintaining communications between the students and the system to deliver the learning contents through VoIP providers such as Skype.
2. Call Originator Software: This software daemon is responsible for scheduling and establishing each call. The schedule management interface is exposed to the content providers by a web service. The functionalities of this software module are enabled by Asterisk Gateway Interface (AGI) script and the Java framework.
3. Call Manager Server: This server is responsible for actual interactions with the students. It defines the logic of each learning session: how audio clips should be played and how it should record the user's response. The logic is implemented using the AGI script. This server exposes its management interface to the administrators using PHP, Apache HTTP server and the Drupal content management system.
4. Database servers: These servers keep track of the students' profiles and audio lectures using MySQL.

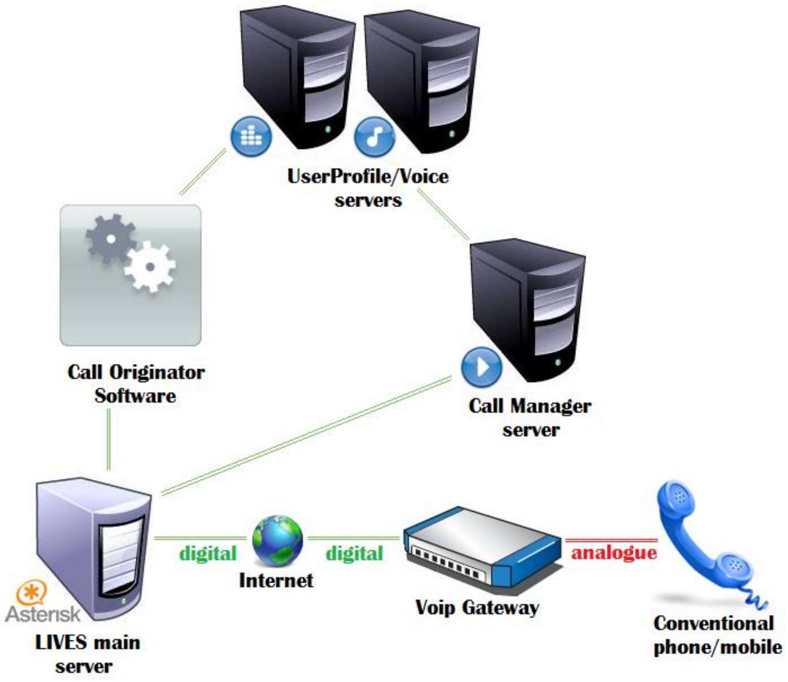


Figure 1 A diagram of the LIVES architecture showing how each component is connected to the others.

5.1.2 LIVESMOBILE Design and Implementation

LIVESMOBILE is a set of extensions to LIVES. The extended components are the Short Message Service (SMS) server to send and receive SMS messages, the Multimedia Messaging Service (MMS) server to send multimedia messages, and the modified Call Originator Software to dispatch MMS. The integration of MMS widens the target audience of LIVES to include the people who have multimedia capable mobile devices. As LIVESGEO allows creation of video contents on a mobile device, it can be used to create video based learning contents in low quality through its mobile client and deliver the contents via LIVESMOBILE as MMS messages. Figure 2 depicts the extended components of LIVESMOBILE:

1. SMS Server: This server processes all incoming and outgoing text messages through the GSM modem attached to the server. The incoming text messages adhere to the syntax

defined by the SMS server to interpret and perform requested functionalities, such as updating user information. It is implemented using open source SMS gateway software called Kannel. Kannel allows a user to define a programmable logic on the GSM modem and enables communication between the server and the modem.

2. MMS Server: This server implements a Multimedia Message Switching Center (MMSC) to send MMS messages directly to a phone circumventing the need to deliver messages via a network service provider. It is implemented using open source MMS gateway software called Mbuni. Mbuni provides a simple interface for management of sending multimedia contents via the GSM modem.

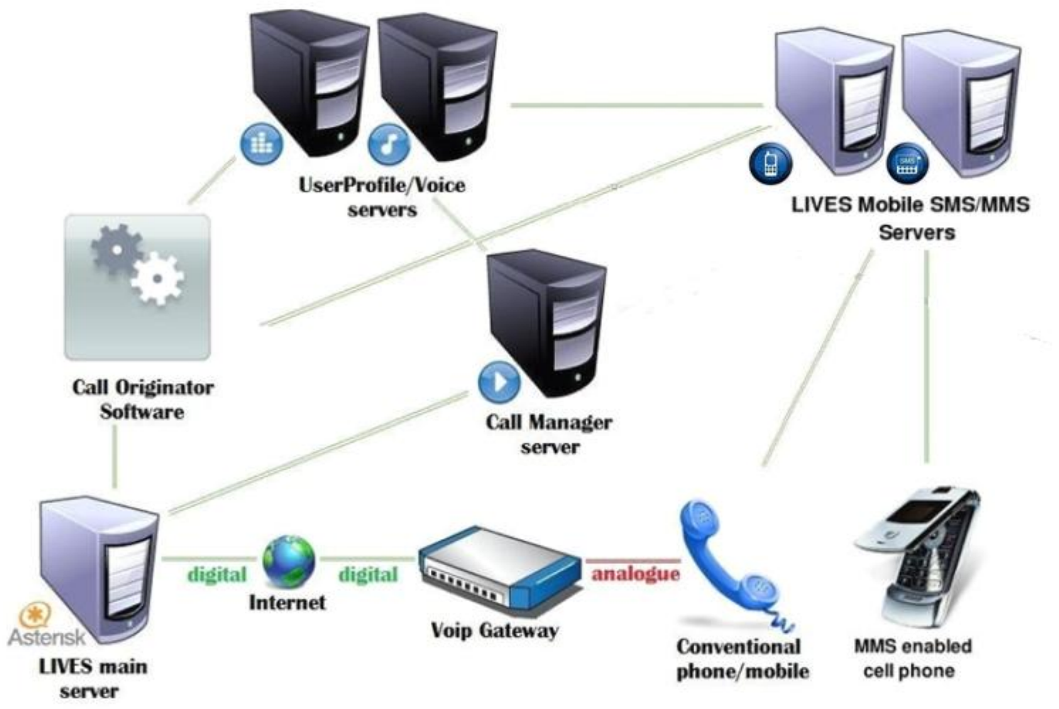


Figure 2 A diagram of the LIVESMOBILE architecture showing how it extends from LIVES.

5.2System Objectives

Several goals were defined for the design and implementation of LIVESGEO. The system aims to overcome the limitations of the previous systems and create a peer driven knowledge sharing system. The system also aims to encompass a wider range of audiences, illiterate and literate populations in developing and developed countries. Therefore, the design and implementation were primarily focused on achieving the following features:

1. Usability: The ability to provide easy and intuitive interfaces to generate, disseminate, and search for geo-tagged multimedia knowledge contents.
2. Extensibility: The ability to improve and extend the predecessors' feature sets without requiring fundamental changes to their design and implementation.
3. Accessibility: The ability to be openly accessible, to anyone and at anyplace without any restriction, to provide open peer driven learning.
4. Interactivity: The ability to encourage the users to engage in the knowledge creation and sharing process.

5.3System Design

Two main concerns of LIVESGEO were how to efficiently manage the multimedia contents and distribute to the users, and how to efficiently distribute the LIVESGEO mobile application to users to drive peer learning. In order to satisfy these two main criteria, the architecture of LIVESGEO adapts to a hybrid of client-server and peer-to-peer models.

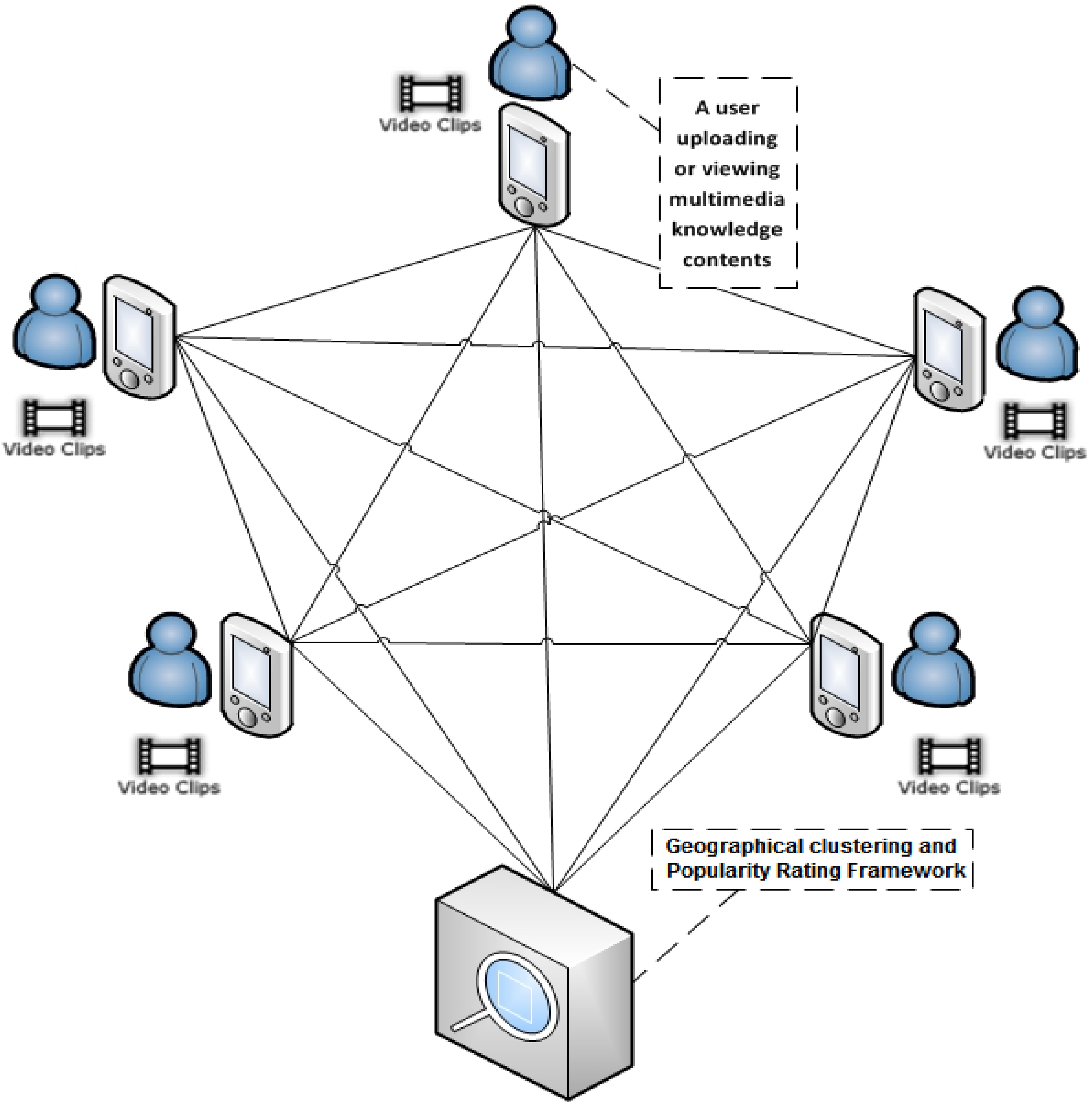

Video Clips
A user uploading or viewing multimedia knowledge contents
Video Clips
Video Clips
Video Clips
Video Clips
Geographical clustering and Popularity Rating Framework

Figure 3 A peer-to-peer approach to peer driven multimedia knowledge content distribution.

The initial architectural consideration of LIVESGEO was a pure peer-to-peer model to distribute and share multimedia knowledge contents between the mobile clients, as depicted by Figure 3.

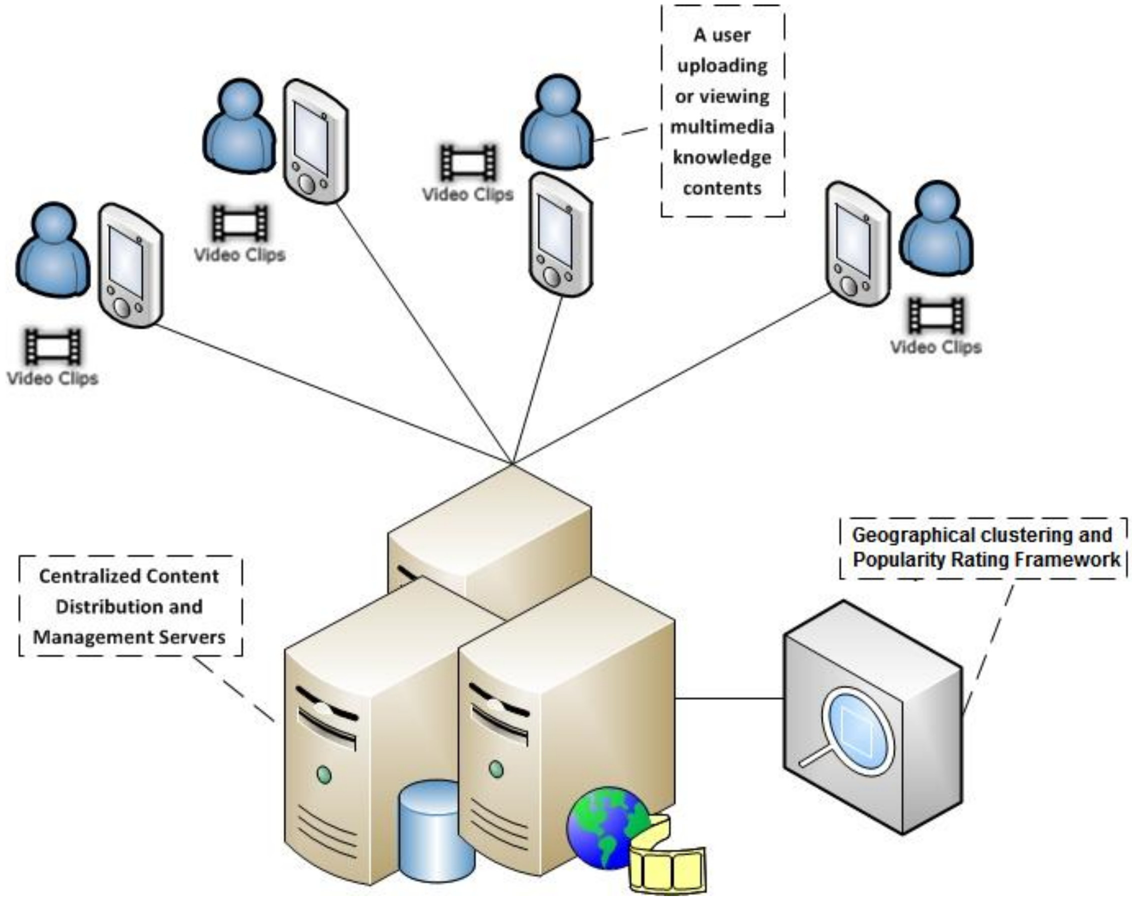

A user uploading or viewing multimedia knowledge contents
Video Clips
Video Clips
Video Clips
Video Clips
Centralized Content Distribution and Management Servers
Geographical clustering and Popularity Rating Framework

Figure 4 A client-server approach to multimedia knowledge content distribution.

However, after putting a considerable amount of time on implementing the peer-to-peer based LIVESGEO during its early stage, we decided to change the direction and adapt to the client-server architecture as depicted by Figure 4. In the client-server based LIVESGEO, the clients will create and upload multimedia contents to the main server. The main server will collaborate with other function-specific servers to store, manage and distribute the contents to the clients. Additional software framework will work in parallel with the servers to data mine and analyze the contents. We highlight the main reasons as to why adapting to the client-server architecture for content management and distribution is advantageous over the peer-to-peer architecture. Below are the inherent physical limitations of mobile devices that make the peer-to-peer architecture unattractive:

1. Insufficient bandwidth: Despite the increasing bandwidth of mobile data networks currently evolving from 3G to 4G, the mobile Internet access is still generally slower than direct cable connections, such as optical fiber. Insufficient bandwidth discourages mobile devices from frequent transmission of high volume of traffic demanding multimedia contents, thus making the centralized server model a more attractive option.
2. Range limit: Mobile devices can keep the connection to the Internet only within the range of either a cellular base tower or a wireless access point. Moreover, it is prone to transmission interferences by various environmental obstacles such as weather, surrounding high rises, and tunnels. This, again, makes the centralize server model a better choice to store and distribute contents.
3. Cost advantages: The data traffic on mobile networks is more costly as mobile data service providers charge relatively higher price on the mobile data traffic.
4. Security standards: The mobile data traffic does not adhere to a particular security protocol standard making itself susceptible to a wider range of security issues. Moreover, it requires technical experts to setup Virtual Private Network (VPN) for mobile devices to enhance security. On the other hand, a centralized server can be protected by more mature operating systems and advanced network security tools, such as firewall, and intrusion detection and prevention systems.
5. Power consumption: Mobile devices must rely on battery power when a direct power connection is unavailable. This limitation is inherent due to its compact size, which often means a relatively short and sporadic connection to the Internet. On the other hand, a centralized server provides a constant connection.

Below are the advantages of using the client-server architecture over the peer-to-peer architecture in terms of design and functionality:

1. Fragmented contents distributed over mobile clients in a peer-to-peer manner is overly complicated to index and search for when needed without providing any extra benefit of using the peer-to-peer over the client-server. Moreover, there is no guarantee that the mobile devices that contain the fragments of the content on demand is turned on and connected to the Internet, making the problem harder. Thus, it is easy to manage the contents at a single point.

2. The concern of a single point of failure (SPOF) can be resolved by various software and hardware SPOF detection and prevention technologies.
3. For the geographical clustering and popularity rating frameworks, Figure 3 and 4, to work efficiently, it requires quick access to the whole geo-tagged multimedia contents. Thus, it is more effective to allow access to the contents at a single point.
4. The contents can be easily reviewed and evaluated for its credibility, security, legality, and quality when it is stored in centralized servers.

Now that we have explained our rationale behind using the client-server model, let's discuss how we decided to distribute the LIVESGEO mobile application to users the most effectively to drive peer learning. In order to do so, the application distribution part of the system adapts to both peer-to-peer and client-server approaches as depicted by Figure 5.

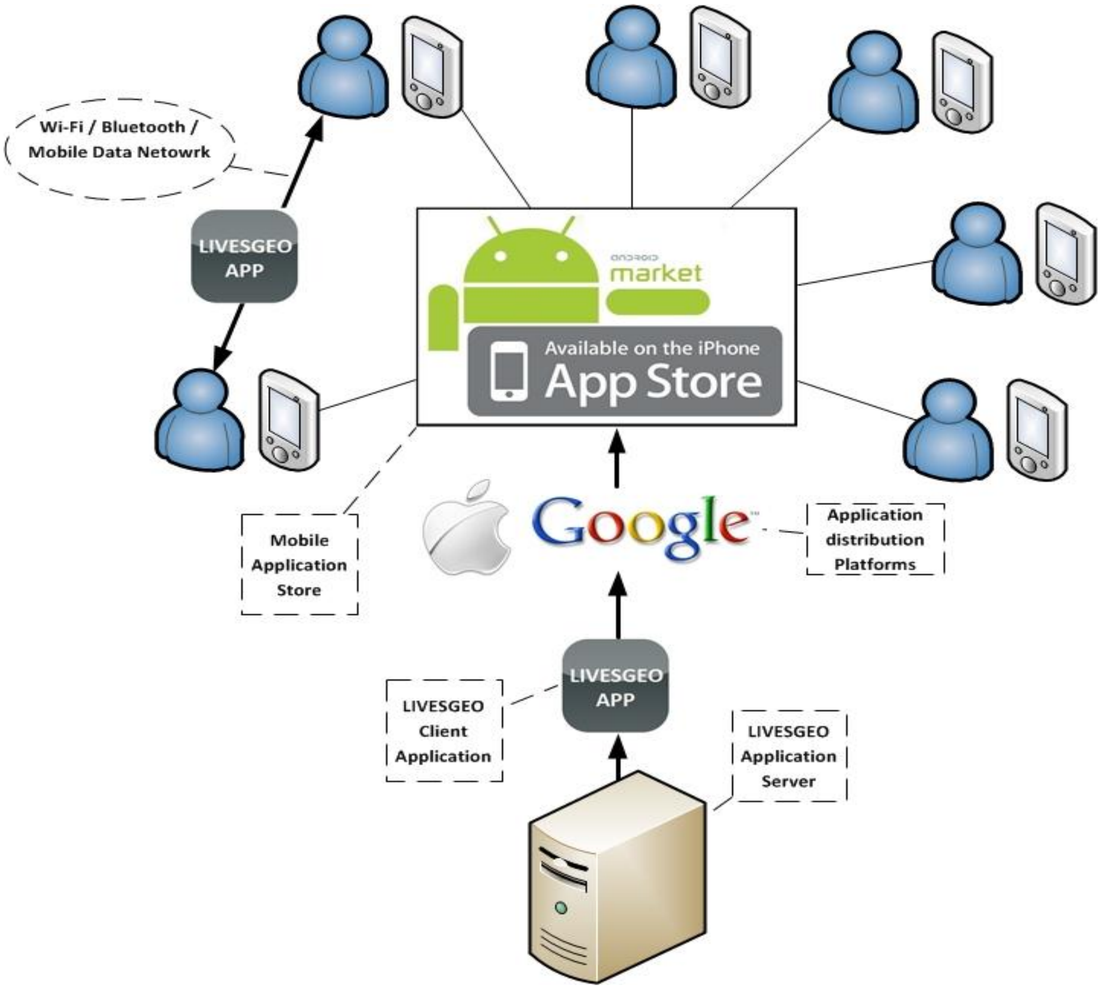

Wi-Fi / Bluetooth / Mobile Data Netowrk
LIVESGEO APP
android market
Available on the iPhone App Store
Google
Mobile Application Store
Application distribution Platforms
LIVESGEO APP
LIVESGEO Client Application
LIVESGEO Application Server

Figure 5 The centralized and peer-to-peer application distribution model. The LIVESGEO application can be submitted and uploaded to application distribution ecosystems or distributed from a user to a user over various communication channels.

The mobile application development environment is setup on the LIVESGEO application server. When the mobile application is developed and updated, it is registered to application distribution service providers, such as Google and Apple. The application is then reviewed by the distribution service providers to ensure the reliability and quality of the application. After the approval, the application can be published through application distribution ecosystems, such as App Store and Application Market. Through these ecosystems, end-users can easily find and install the LIVESGEO mobile application. Moreover, the users will automatically be notified if the application needs to be reinstalled or updated through these ecosystems. Distributing the LIVESGEO application through the centralized mobile application stores provides us the most convenient way to disseminate and update our mobile application. Another major benefit of distributing applications through these ecosystems is that we can monetize the application traffic to produce revenue, by inserting paid advertisements on the user interface of the application. This revenue can be used to fuel the development of the system or reward the peers providing highly rated contents for others.

As an alternative to the centralized distribution approach, the LIVESGEO mobile client is implemented in a way that it allows the application itself to be transferred from a peer to a peer. The user of LIVESGEO can send the application over the Internet, using popular services like email and instance message, Bluetooth, and SMS, over a mobile data network. Peer-to-peer distribution may seem unnecessary, at first glance, as downloading an application from application stores is far more intuitive and easy. However, mobile application stores are not necessarily available on every mobile device due to various technical and business reasons. For example, an eBook reader may be running Android over a mobile data network or Wi-Fi, but it may not have an application store because eBook is, by design, to be primarily used as a reading device. Moreover, the vendors of eBook device may limit the application store functionalities as they want to provide their own proprietary application store to sell only their own electronic contents to maximize their revenue. Therefore, peer-to-peer application distribution will accommodate a broader range of users with different mobile devices.

Figure 6 depicts the overall architecture of LIVESGEO. The LIVESGEO application server publishes the mobile application through application ecosystems. As explained above, the application can also be shared in a peer-to-peer manner. The users create multimedia knowledge contents from the LIVESGEO mobile client and upload it to the LIVESGEO application server over a cellular data network or Wi-Fi. The users will be able to rate the contents they like, therefore every content will have rating meta-data associated with itself. The LIVESGEO application server will coordinate the management of the multimedia contents between the geo-location content server and the multimedia server. The geographical clustering and popularity rating frameworks will extract the geo-tagged content datasets from the aforementioned servers and provide services to the calling applications and users. The LIVESGEO application server can be the access point to provide a web-based or application-based administrator interface to the administrators if the contents and the registered users need to be managed at some level.

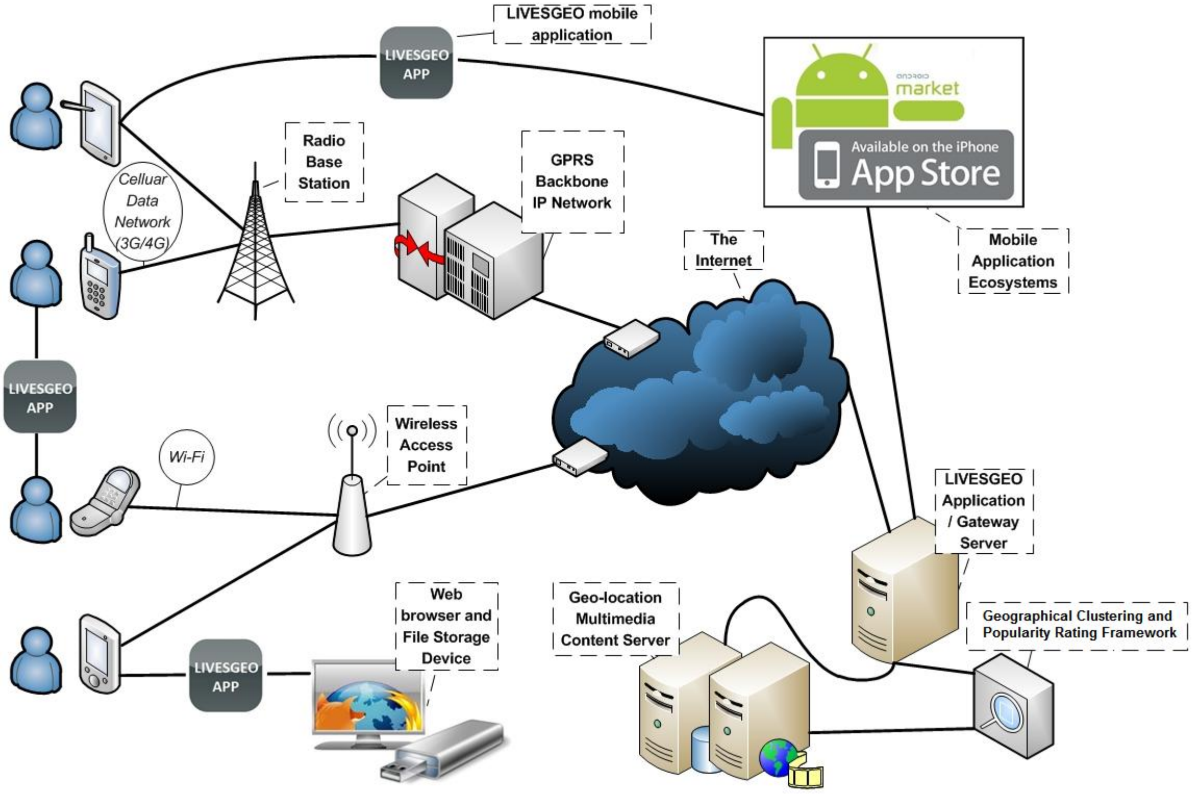

LIVESGEO mobile application
LIVESGEO APP
android market
Available on the iPhone
App Store
Radio Base Station
GPRS Backbone IP Network
Celluar Data Network (3G/4G)
The Internet
Mobile Application Ecosystems
LIVESGEO APP
Wi-Fi
Wireless Access Point
LIVESGEO Application / Gateway Server
Geographical Clustering and Popularity Rating Framework
Web browser and File Storage Device
Geo-location Multimedia Content Server
LIVESGEO APP

Figure 6 The overall architecture of LIVESGEO.

5.4Mobile Client Implementation

The design of the LIVESGEO mobile application is focused on making the application simple and intuitive for the users to create, upload, search, view, and rate knowledge contents effortlessly. This is achieved by taking full advantages of software libraries and hardware features of smart-phones. The LIVESGEO mobile application aims to provide:

1. A simple way to quickly create a multimedia knowledge content on the spot.

2. An intuitive way to upload a multimedia knowledge content to the repository seamlessly over different communication mediums.
3. An entertaining way to access the knowledge contents to maximize the user's engagement in content sharing.

The simplicity and intuitiveness of the application is crucially important as the physical limitations, described in 5.3, of smart-phones makes itself hard to develop complex and sophisticated ways to create, share, and view knowledge contents. In order to make the application as simple and intuitive as possible, at the same time providing all the sophisticated features, we developed an android application on the Android 2nd generation platform using its network/multimedia/geo-location API, SQLite database framework, Google's Map API and external Apache HttpClient libraries.

Figure 7 depicts a prototype implementation of the LIVESGEO mobile application and how it provides simple and intuitive ways for the users to create, upload, search, view, and rate knowledge contents. A user will be able to create one's own account and sign into the system and be prompted with the main screen. The upload menu will allow the user to take a video or audio clip using the phone's built-in camera and microphone and multimedia API to create a multimedia knowledge content. The user will be able to define the title of the content, select an appropriate category, and upload the content to the LIVESGEO repository. The content will be broken down into multiple messages and posted to the server as multi part messages using extra Apache HttpClient libraries. We noticed that Android does not support the MultipartEntity Interface by default, which is used to post multipart messages to the web server via the HTTP POST request. In order to resolve this problem, external HttpClient library[2] from Apache, which has MultipartEntity implemented, was compiled along with the application and imported from the application's class that is responsible for making network connections. A newly created content will automatically be geo-tagged through geo-location API. The search menu will allow the users to navigate a virtual map, and search and view geo-tagged multimedia knowledge contents in vicinity. The user's current location is automatically detected by using geo-location API. The location can be estimated using either coarse location, which is calculated using nearby

[2] Although, Android claims to use a unique Virtual Machine called Dalvik which is different from Java Virtual Machine (JVM), an external Apache library that was written for JVM was compiled and interpreted without hindering the library's functionalities.

wireless access points or cellular base towers, or fine location, which is calculated using built-in GPS. The contents will be displayed on the map as a layer of icons on top of the virtual map using the Google Map Overlay API. A user will be able to selectively filter out the contents on the map by selecting the categories that the user wants to remove from the map. A user will be able to view and rate the selected content. The rating information will be sent back to LIVESGEO's back-end servers. The recommendation menu will provide interesting information about the contents in the user's vicinity. The share app menu will allow the application to be distributed to other peers over Bluetooth or Wi-Fi.

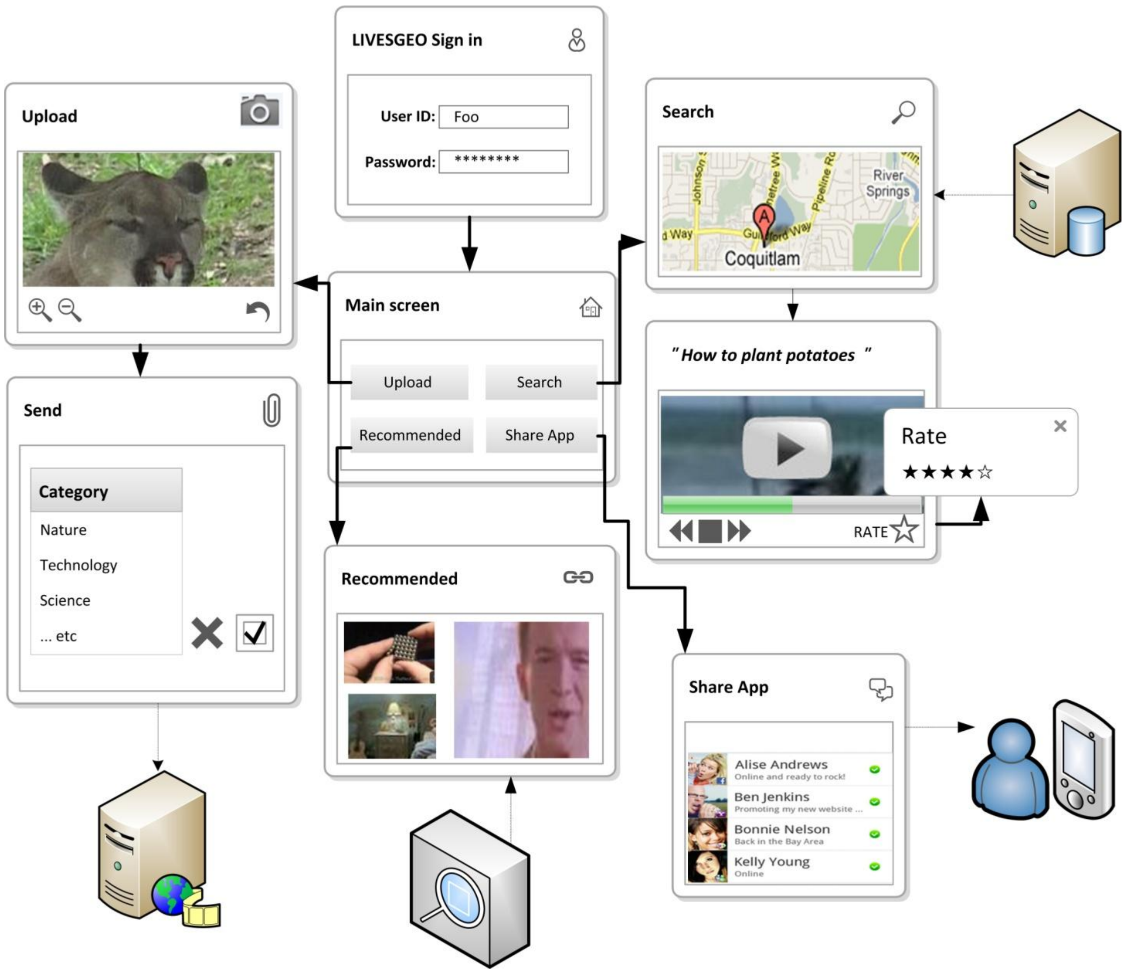


Figure 7 A diagram of the LIVESGEO mobile application. Its main goal is to provide simple and intuitive ways for the user to create, share, view, and rate knowledge contents.

5.5Server Side Implementation

LIVESGEO's back-end servers are designed to process the requests from the mobile clients and store and manage the multimedia contents. The main functionalities are:

1. Receive the multimedia contents from the users and stream the multimedia contents to the users.
2. Manage the video and audio clips in the multimedia database.
3. Manage the user profile and rating, and geographical information.
4. Expose a necessary portion of the database for other modules and frameworks to be built upon and leverage the geo-tagged multimedia contents.

The back-end implementation was relatively trivial as there are a number of popular software available that we can take advantages of to implement the aforementioned functionalities. The server side logic was implemented using Servlet, which is a Java network framework, to extend the capabilities of the servers to take programmed actions upon the mobile client's HTTP requests and also to access the multimedia and profile databases within the programmed logic. Apache Tomcat was used to run a HTTP server to receive multipart video or audio contents, stream multimedia files, and serve as a container to run server side logic, namely Servlet. MySQL, was used to manage the multimedia contents, geo-location information, user profiles, and content rating datasets. Java Databse Connectivity API was used to act as a bridge between servlets and MySQL to allow database access from the Apache server to the databases so that the server can provide dynamic contents for each client request.

6 Content Popularity Approximation

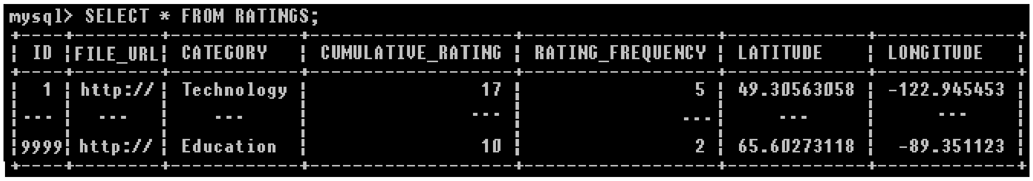
LIVESGEO enables the users to rate contents. The ratings of contents provide a numerous ways to analyze interesting trends. For instance, the system can provide the top ten contents of each category in the user's vicinity as the system has the information regarding the ratings of contents, the user's current location, and the geographical coordinates of all the contents in the system. Moreover, as explained in 4.1, it is challenging to enforce high quality of contents while allowing for complete freedom. The ratings of contents allow the system to conform to the idea of enforcing community introspection by rewarding the users with more informative and high quality contents. In order to do so, we need a way to measure the popularity of a content relative to all other contents of interest. In other words, if we want to represent a particular content's popularity within the city of Vancouver, we need to measure and represent the popularity of this content relative to the popularity of all the other contents posted within Vancouver.

In this chapter, we look at how the popularity of a content can be best approximated using the content's own rating and the ratings of other contents. We describe the logic behind each measurement we have developed and provide description of how the optimal measurement needed by the system is obtained. We present the evaluations of our novel measurements against popular movie content rating providers to justify the performance of our measurements.

6.1Cumulative Rating and Rating Frequency

The LIVESGEO server keeps track of two important attributes for each geo-tagged multimedia content instance in the database, which are the cumulative rating and the rating frequency. These attributes are used to measure the popularity of a content. Figure 8 depicts the multimedia content table stored in the LIVESGEO server. The table represents a set of geo-tagged multimedia contents. Each row represents a multimedia content instance. An instance is newly created when a user uploads a new content and is updated when its rating values are changed. Each instance contains the attributes: FILE_URL, the address of the corresponding media file in the file server, CATEGORY, the category of the content that the creator has defined, CUMULATIVE_RATING, the sum of all ratings, RATING_FREQUENCY, the number of times the content was rated, and

LATITUDE and LONGITUDE, the geographical coordinates of the content. For the sake of simplicity, we assume that rating value ranges from 1 to 5, representing worst to best respectively.



```
mysql> SELECT * FROM RATINGS;
+----+--------+------------+-------------------+------------------+-------------+-------------+
| ID |FILE_URL| CATEGORY   | CUMULATIVE_RATING | RATING_FREQUENCY | LATITUDE    | LONGITUDE   |
+----+--------+------------+-------------------+------------------+-------------+-------------+
|  1 | http://| Technology |                17 |                5 | 49.30563058 | -122.945453 |
|--- |  ---   |    ---     |               --- |               ---|     ---     |     ---     |
|9999| http://| Education  |                10 |                2 | 65.60273118 |  -89.351123 |
+----+--------+------------+-------------------+------------------+-------------+-------------+
```

Figure 8 A table of multimedia contents. Each row represents a multimedia content instance. The CUMULATIVE_RATING attribute represents the sum of all ratings on the content. The RATING_FREQUENCY attribute represents the number of times the content was rated. The LATITUDE and LONGITUDE attributes represent the geographical coordinates of the content.

6.2Measurements

In order to measure the popularity, we focus on developing a rating measurement which fairly represents the reputation and the quality of a content relative to others using the cumulative rating and rating frequency values. We will start with a trivial measurement and develop into to a pragmatic measurement explaining the rationale behind each development.

6.2.1 Raw Rating Measurement

The most intuitive and trivial measurement can be designed by utilizing the given rating values of a content, which are the cumulative rating and rating frequency values. An arithmetic mean can be calculated for each content using its cumulative rating and rating frequency values and used to represent the raw rating of the content. The mathematical representation of the raw rating measurement is:

$$raw\_r = \frac{\mathrm{cr}}{\mathrm{rf}} \tag{6.1}$$

where raw_r represents the popularity of the content, cr represents the cumulative sum of all the ratings of the content, and rf represents the rating frequency of the content. It is essentially the same as Equation 6.2, where n denotes the total number of the ratings and x_i denotes each rating value.

$$raw\_r = \frac{1}{\mathrm{n}} \cdot \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i} \tag{6.2}$$

Although the raw rating measurement using the arithmetic mean well represents the absolute popularity of the content itself, it fails to represent the popularity of the content among all the other contents. In other words, it does not take the ratings of all the other contents into account to represent itself relative to the others. For instance, in Figure 9, the content A's popularity is $(4 + 4 + 4 + 3)/4 = 3.75$ and the content B's popularity is $4\ /\ 1 = 4$. As illustrated by Figure 9, the raw rating measurement fails to represent the true popularity as it scores the content B higher. The content A has been rated more frequently, which indicates that it has a higher reputation among the users, and rated as high as the content B except for the one rating of 3.The single rating of 3 of the content A is overrepresented and does not contribute as a fair comparison value. The raw rating measurement fails to capture the underlying nature that the higher rating frequency indicates the higher reputation of the content and the higher confidence level of its ratings.

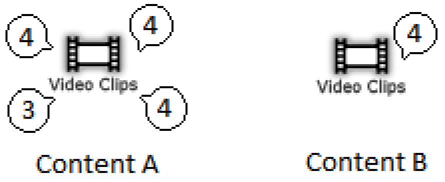


Figure 9 Content A with three ratings of 4 and one rating of 3 and Content B with one rating of 4.

6.2.2 Reputation Weighted Measurement

The raw rating measurement demonstrated that the rating frequency, which represents a reputation, must be taken into account to represent the popularity of a content relative to the others. Also, a higher rating frequency indicates a higher confidence level, which we will define as the likelihood that the true relative popularity is represented by the cumulative rating of the content. In order to incorporate the concept of a reputation, which is expressed by the rating frequency, to our measurement framework, Equation 6.1 is modified to penalize a content if its rating frequency falls below the threshold, which we set to the average rating frequency of all contents. A content is penalized or rewarded by the reputation weight variable. The mathematical representation of the threshold is:

$$t_{rep} = \frac{1}{\mathrm{n}} \cdot \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{rf}_{\mathrm{i}} \tag{6.3}$$

where $\mathrm{t}_{\mathrm{rep}}$ denotes the threshold, n denotes the number of contents, and rf_{i} denotes the rating frequency of each content. The threshold is simply the arithmetic mean of all the rating frequency values of the contents. The mathematical representation of the reputation weight is:

$$rw = \left\{ \left(\frac{\mathrm{rf}}{\mathrm{t_{rep}}} \right) \text{or 1, if greater than 1} \right\} \cdot \text{ max \_r} \tag{6.4}$$

where rw denotes the reputation weight and max_r denotes the maximum rating value that can be given to a content. The reputation weight is a weighted value to penalize a content if its rating frequency is comparably lower than the average rating frequency of all the contents. Assuming that max_r is 5, rw would range from $0 < rw \leq 5$, ensuring that rw never becomes zero. The mathematical representation of the reputation weighted measurement is:

$$rw\_r = \frac{(\mathrm{raw\_r} + \mathrm{rw})}{2} \tag{6.5}$$

where rw_r denotes the reputation weighted rating. Equation 6.5 will penalize the rating of a content that has not been rated frequently or which its rating confidence level is low relative to the others.

The reputation weighted measurement does a better job on fairly representing the popularity of a content taking the concept of a reputation as a factor of the popularity. However, as we were testing the reputation weighted measurement, we discovered that the reputation weight is often overrepresented, as demonstrated in Table 1.

Content ID	**Ratings**	**Raw rating measurement**	**Reputation weighted measurement**
A	1,1,1,1,1,1,1,1,1,1	1	3
B	5,4,1	3.33	2.82

Table 1 The result of applying the raw rating measurement and the reputation weighted measurement on the content A and B. The raw rating measurement captures the true popularity whereas the reputation weighted measurement fails to do so. The reputation weighted measurement scores the content A higher than the content B.

The reputation weighted measurement often gives an advantage to the ones with lower ratings but a higher rating frequency just because of the sole fact that it has been rated more frequently despite its low rating score quality. Adjustments to the reputation weight variable, rw, did not resolve the fundamental problem. The reputation weighted measurement could not correctly capture the popularity when there was a significant difference between the rating frequencies of two contents. Due to this problem, the reputation weighted measurement will not give a fair chance to a highly rated content that has been only recently uploaded, which inevitably has a lower rating frequency than previously uploaded contents, regardless of its quality.

6.2.3 Quality Weighted Measurement

To overcome the limitation of the reputation weighted measurement, we decided to alter the measurement framework to also penalize the low rating score separately from the reputation using the quality weight variable. A content is penalized when its average rating score, which is calculated by dividing its cumulative rating value over its rating frequency value, falls below the threshold, which is the arithmetic mean of the average rating scores of all the contents. The quality weighted measurement will penalize a content that has been rated a lot but its rating quality is low. The mathematical representation of the threshold is:

$$t_{qlt} = \frac{1}{\mathrm{n}} \cdot \sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{\mathrm{cr_i}}{\mathrm{rf_i}} \tag{6.6}$$

where $\mathrm{t_{qlt}}$ denotes the quality threshold, $\mathrm{cr_i}$ denoted the cumulative rating value of each content, and $\mathrm{rf_i}$ denotes the rating frequency value of each content. $\mathrm{t_{qlt}}$ represents the average of the set of average rating scores from every content. The mathematical representation of the quality weight is:

$$qw = \left\{ \left(\frac{\mathrm{cr\_i/rf\_i}}{\mathrm{t_{qlt}}} \right) \text{ or } 1, \text{if greater than } 1 \right\} \cdot \max \_\mathrm{r} \tag{6.7}$$

where qw denotes the quality weight and $\mathrm{max\_r}$ denotes the maximum rating value that a content can be given. The quality weight is a weighted value to penalize a content having its average rating score comparably lower than the other contents' average rating scores. Assuming that $\mathrm{max\_r}$ is 5, qw would range from$0 < \mathrm{qw} \leq 5$, ensuring that qw never becomes zero. The mathematical representation of the quality weighted measurement is:

$$qw\_r = \frac{(\mathrm{raw\_r} + \mathrm{rw} + \mathrm{qw})}{3} \tag{6.8}$$

where $\mathrm{qw\_r}$ denotes the quality weighted rating, $\mathrm{raw\_r}$ denotes the raw rating value, rw denotes the reputation weight, and qw denotes the quality weight.

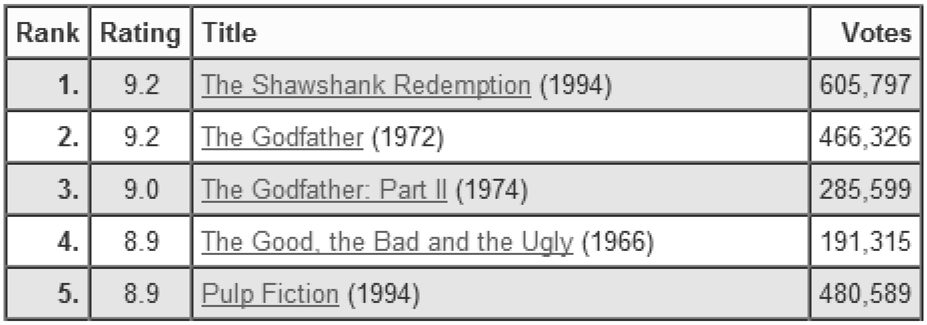
As shown in Table 2, the quality weighted measurement ameliorates the limitation of the reputation weighted measurement and more fairly represents the relative popularity of contents taking the concept of both reputation and quality into account.

Content ID	Ratings	Raw rating measurement	Reputation weighted measurement	Quality weighted measurement
A	1,1,1,1,1,1,1,1,1,1	1	3	2.77
B	5,4,1	3.33	2.82	3.54

Table 2 The result of applying the raw rating measurement, the reputation weighted measurement, and the quality weighted measurement on contents A and B. Note that the quality weighted measurement represents the popularity of the contents more fairly.

6.3Evaluation

In this section, we present the results and analyses of the evaluation of our popularity approximation framework to justify its performance and accuracy. We compare our measurements against the movie ranking measurement of a popular online movie database, Internet Movie Database (IMDb). In order to do so, we used the movie ranking from IMDb's top 250 movie ranking data, shown in Figure 10, as the reference point to draw insightful assessment of our measurements. In the following subsections, we describe how the experimental dataset has been setup and the criteria we were using to compare the performance of our measurements against IMDb.



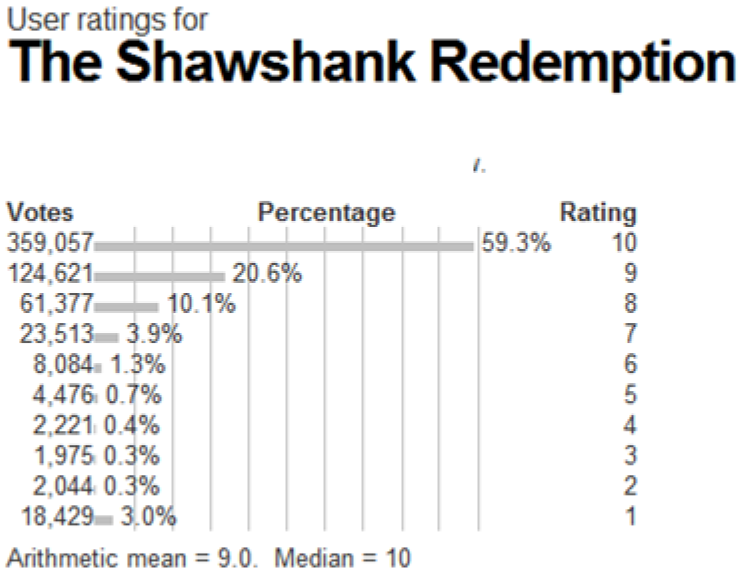
Rank	Rating	Title	Votes
1.	9.2	The Shawshank Redemption (1994)	605,797
2.	9.2	The Godfather (1972)	466,326
3.	9.0	The Godfather: Part II (1974)	285,599
4.	8.9	The Good, the Bad and the Ugly (1966)	191,315
5.	8.9	Pulp Fiction (1994)	480,589

Figure 10 A portion of IMDb's top 250 movie ranking.

6.3.1 Experimental Setup

IMDb is an online database of information related to movies. It is one of the most complete and popular movie databases on the Internet. IMDb offers a rating scale which allows the users to rate a movie from one to ten, representing 'awful' to 'excellent' respectively. IMDb's rating system rates and ranks movies using its own rating framework. It applies various filters and weighted rating variables, which are not disclosed to the public, on the raw user ratings to reduce biased voting and represent the ranking of each movie as accurately as possible relative to other movies. Although, the algorithm of their rating framework is not disclosed to the public, the movie ranking of IMDb serves as a desirable reference point to compare the performance of our own framework.

As IMDb provides the demographic breakdown of the exact number of voters per each rating score of a movie, shown in Figure 11, it was possible to extract the cumulative rating value and the rating frequency value of a movie listed in top 250. Figure 11 shows how the cumulative rating value and the rating frequency value were calculated using the information provided by IMDb. 40 samples of the ranked movies from IMDb's top 250 were selected as a reference group and this dataset was fed into our measurements to produce the popularity rankings. Figure 12 shows an instance of the experimental dataset fed into our measurements.



User ratings for

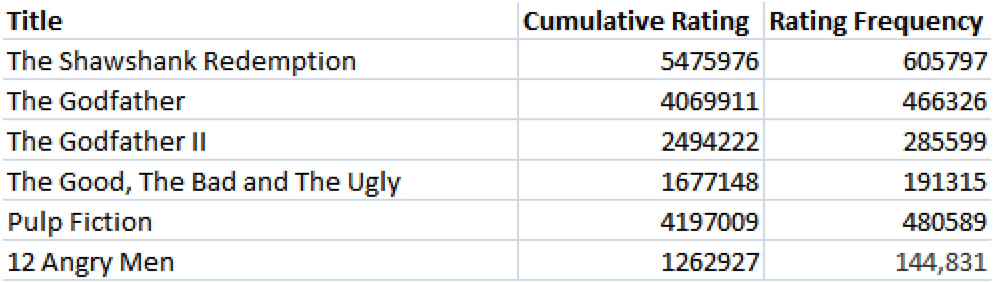
The Shawshank Redemption

Votes	Percentage	Rating
359,057	59.3%	10
124,621	20.6%	9
61,377	10.1%	8
23,513	3.9%	7
8,084	1.3%	6
4,476	0.7%	5
2,221	0.4%	4
1,975	0.3%	3
2,044	0.3%	2
18,429	3.0%	1

Arithmetic mean = 9.0. Median = 10

Rating frequency	359,057+124,621+61,377+23,513+8,084 +4,476+2,221+1,975+2,044+18,429	605,797
Cumulative rating	359,057(10)+124,621(9)+61,377(8)+23,513(7)+8,084(6) +4,476(5)+2,221(4)+1,975(3)+2,044(2)+18,429(1)	5,475,976

Figure 11 The demographic break down of the user ratings on a movie from IMDb and the cumulative rating and rating frequency values calculated based on the information.



Title	Cumulative Rating	Rating Frequency
The Shawshank Redemption	5475976	605797
The Godfather	4069911	466326
The Godfather II	2494222	285599
The Good, The Bad and The Ugly	1677148	191315
Pulp Fiction	4197009	480589
12 Angry Men	1262927	144,831

Figure 12 An instance of the experimental dataset fed into our measurements to produce the movie rankings of these data based on their popularity.

6.3.2 Evaluation Metric

In order to compare the performance of our measurements against IMDb, the ranking deviation of each movie between the ranking of IMDb and the rankings produced by our popularity measurements was used as an evaluation metric. For instance, if the move 'The Matrix' was ranked 5 in IMDb and 9 in our measurement, the deviation would be 4. Lower deviations would indicate positive performance of our measurements relative to IMDb.

6.3.3 Results

Figure 13 plots the ranking deviations between the ranking of IMDb and the rankings produced by the raw rating measurement, the reputation weighted measurement, and the reputation and quality compound weighted measurement on the same movie dataset. All of the measurements show a similar trend to the reference ranking of IMDb. The average absolute deviations, which measure the general deviation tendency of the samples from the reference point, for the raw rating, the reputation weighted, and the reputation and quality compound weighted were 6.5, 5.74, and 4.85 respectively. The result indicates that the combination of reputation and quality weighted measurements outperforms the raw rating measurement and the reputation weighted only measurement, in terms of their performance against IMDb reference point. In other words, our reputation and quality compound weighted measurement, on average, ranked a given movie 4.85 higher or lower than the ranking of IMDb. This does not necessarily mean that our measurement performs inferior to IMDb as IMDB is only used as a reference point to gain an insight about our measurements.

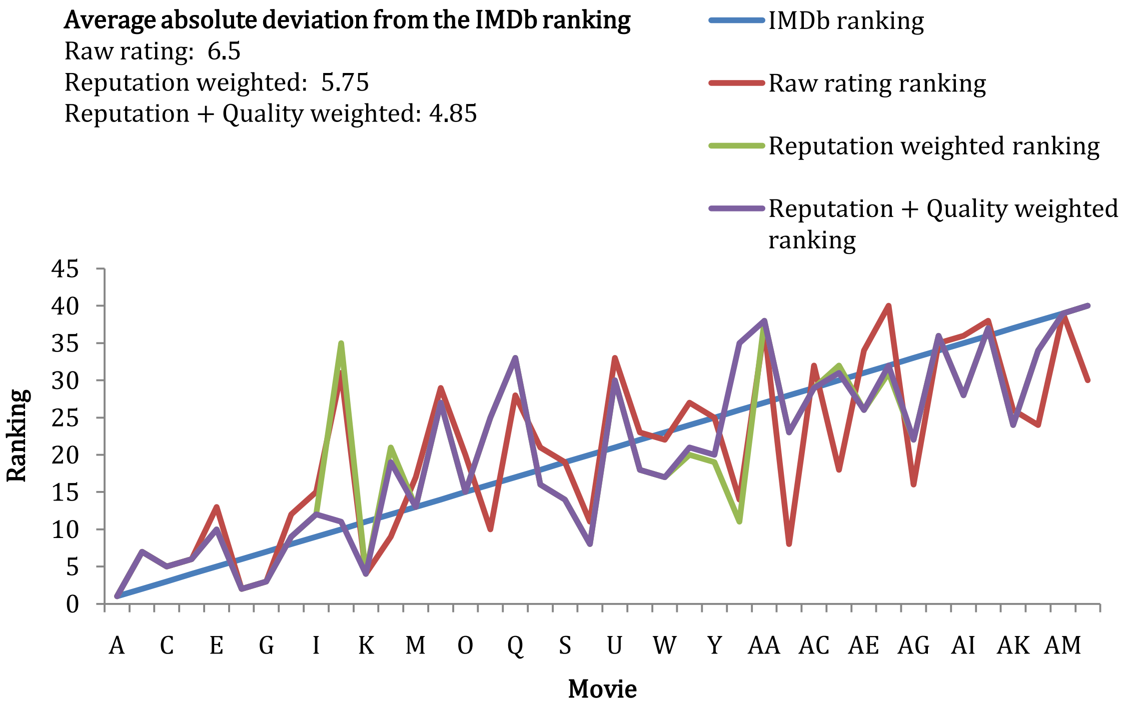


Figure 13 The figure shows the plots of the rankings produced by the raw rating measurement, the reputation weighted measurement, and the reputation and quality compound weighted measurement compared against the IMDb ranking. The absolute deviations from the IMDb ranking for the raw rating ranking, the reputation weighted ranking, and the reputation and quality compound weighted ranking are 6.5, 5.75 and 4.85 respectively. Notice how the reputation and quality compound weighted measurement results in a lower deviation from the IMDb reference point.

Figure 14 plots the rating deviations between the movie rating of IMDb and the popularity ratings produced by the raw rating, the reputation weighted, and the reputation and quality compound weighted on the same movie dataset. The purpose of this comparison is to evaluate the performance of our measurements based on the movie rating score rather than the movie ranking. The absolute deviations from the IMDb movie rating for the reputation weighted rating and the reputation and quality compound weighted rating are 2.09 and 1.91 respectively. Notice how the reputation and quality compound weighted measurement results in a lower deviation from the IMDb reference point rating. The result indicates that the reputation and quality compound weighted measurement outperforms the reputation only measurement producing the trend that more closely resembles the trend of the IMDb movie rating. Our reputation and quality compound weighted measurement produced rating score that is 1.21 higher or lower than the IMDb movie rating score, on average, where the rating score ranges from 0 to 10.

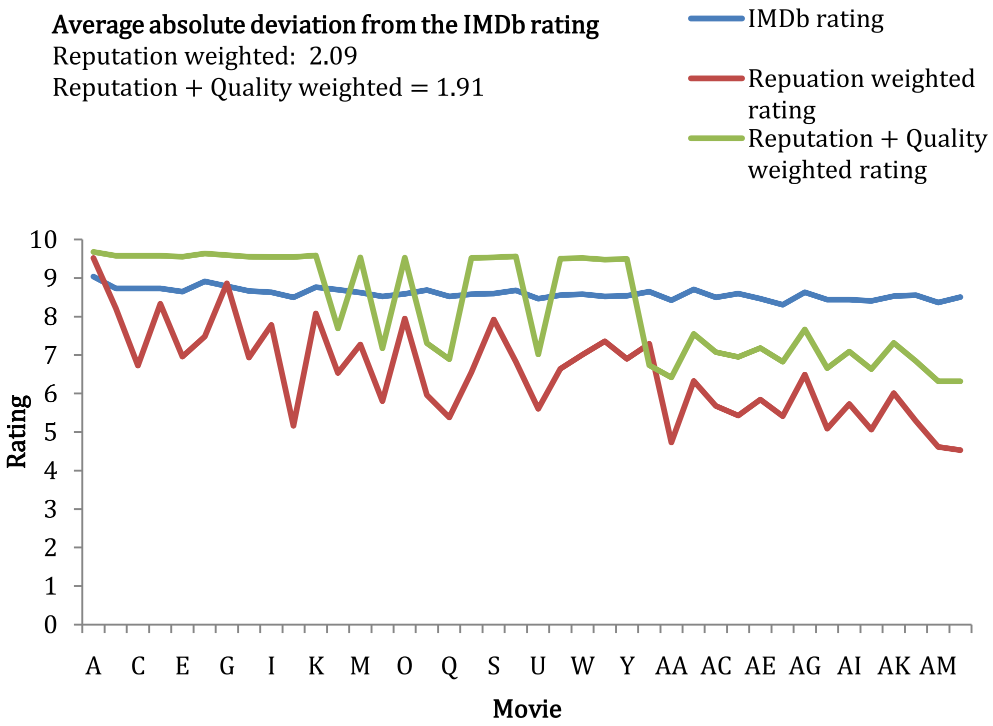


Figure 14 The figure shows the plots of the ratings, produced by the reputation weighted measurement and the reputation and quality compound weighted measurement, compared against the IMDb rating. The absolute deviations from the IMDb rating for the reputation weighted rating, and the reputation and quality compound weighted rating are 2.09 and 1.91 respectively. Notice how the reputation and quality compound weighted measurement results in a lower deviation from the IMDb reference point.

An important fact in here is that our reputation and quality compound measurement was able to grasp the ranking and popularity rating patterns from the raw user rating data and express the trend that closely resembled the IMDb ranking and popularity rating, which is one of the most visited and resourceful movie information databases. However, the difference between the reputation only weighted measurement and the reputation and quality compound weighted measurement was less significant than we had expected. This performance indifference is suspected to be the result of the significantly high and stabilized values of the ratings and rating frequencies of the movies on IMDb. In order to investigate further, another experiment was done, using the same methodology, on another famous movie ranking site with relatively new box-office movies which show large fluctuations in their values of the ratings and rating frequencies.

6.4Additional Evaluation

As the critics' ratings are a good indication of the reputation and quality of a movie, we have selected the top seven currently playing box-office movies and produced the popularity ratings to compare with the critics' ratings given by a famous movie critic site rottentomatoes.com on the same movie dataset. The results are shown in Table 3 and Figure 15.

Title	Rottentomatoes critics rating	Reputation weighted rating	Reputation weighted deviation from critics	Reputation + Quality weighted rating	Reputation + Quality weighted deviation from critics
Midnight in Paris	4.60	3.09	1.51	4.72	0.12
X-men: First Class	4.35	4.55	0.20	4.42	0.07
Kung Fu Panda	4.15	4.50	0.35	4.21	0.06
Super 8	4.10	4.68	0.58	4.52	0.42

Title	Rottentomatoes critics rating	Reputation weighted rating	Reputation weighted deviation from critics	Reputation + Quality weighted rating	Reputation + Quality weighted deviation from critics
Mr. Popper's Penguins	2.30	2.64	0.34	2.53	0.23
Green Lantern	1.30	4.25	2.95	3.25	1.95
Judy Moody	0.75	1.85	1.10	1.02	0.27

Table 3 The popularity ratings, produced by the reputation weighted measurement and the reputation and quality compound weighted measurement, compared against the critics' ratings. Notice how the reputation and quality compound weighted ratings show lower deviations compared to the reputation only weighted ratings.

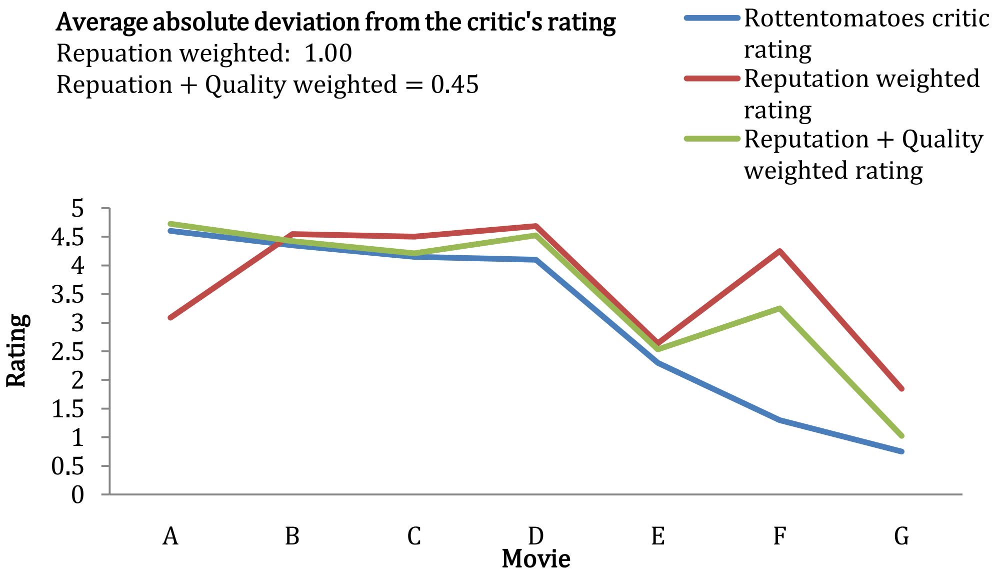


Figure 15 The figure shows the plots of the ratings, produced by the reputation weighted measurement and the reputation and quality compound weighted measurement, compared against the critics' ratings. Notice how the reputation and

quality compound weighted rating more closely resembles the trend of the critics' rating.

Table 3 and Figure 15 show that our measurements were able to generate the ratings that generally agree to the critics' ratings. Our plot does not necessarily have to be the same as the plot of the critics, as sometimes the critics' ratings can be biased and drastically different from the movie's true popularity. Although, both the reputations only weighted measurement and the reputation and quality compound weighted measurement produce the similar trends, the later was able to capture the rating that the former had missed. Moreover, the average absolute deviation of the reputation and quality compound weighted measurement was significantly lower than the reputation only weighted measurement showing 0.45 and 1.00 respectively. The results, again, support the fact that the reputation and quality compound weighted measurement is better capable of grasping the underlying popularity of contents.

7 Spatial Content Clustering

In this chapter, we explore the applicability and usefulness of the geo-tagged contents accumulated by LIVESGEO. To demonstrate this, we abstract the system into the spatial content cluster framework, which acts as an abstract layer between reverse-geocoded contents and applications to offer the multimedia contents as a form of a geographical cluster. This framework can provide services that allow unique applications to be built upon the system, such as a location specific content analysis tool to analyze the quality of the contents within a given geographical zone. We provide examples of such applications that utilize the framework. In order to justify the use of reverse-geocoding to cluster contents geographically, we provide empirical analyses of reverse-geocoding by making novel use of two popular mathematical models in the field of cartography, namely the Haversine and Vincenty formulae.

7.1Classic Clustering Techniques

As each multimedia content instance can be represented as a specific point on a map by latitude and longitude coordinates, the most intuitive choice of an algorithm to geographically group these instances would be a popular data mining technique called clustering. Therefore, we first looked at how a clustering algorithm can be used as the engine to geographically cluster the geo-tagged multimedia contents of LIVESGEO. The distance measurement is crucially important to clustering algorithms to measure the distance between the items in order to cluster them. As the geographical distance between geo-tagged contents is symmetric, the distance measurement is very intuitive, we can simply use the Manhattan distance or the Euclidean distance on the latitude and longitude coordinates between two data instances to calculate their proximity. We can assign each city center's geographical coordinates as a centroid point and perform agglomerate clustering using K-means or its derivative clustering methods as new instances accumulate. Although, the idea was sound, we quickly realized that clustering is rather an inefficient choice for the following reasons:

1. When a user is given an option to define n different geographical ranges of interest, there must be n number of the same clusters only with different range coverage. For example, if a user is interested in information about the contents within the street and city ranges from the user's current location, the framework must have the street-range and city-range clusters pre-clustered and ready for the user.
2. Ambiguity rises when an instance falls into more than one cluster.

Figure 16 provides a visualization of these problems. As the figure depicts, the framework needs to pre-compute differently sized clusters on the same centroid in order to provide n clusters with different ranges. The right most clusters in the figure depict the problem of ambiguity rising from overlapping clusters.

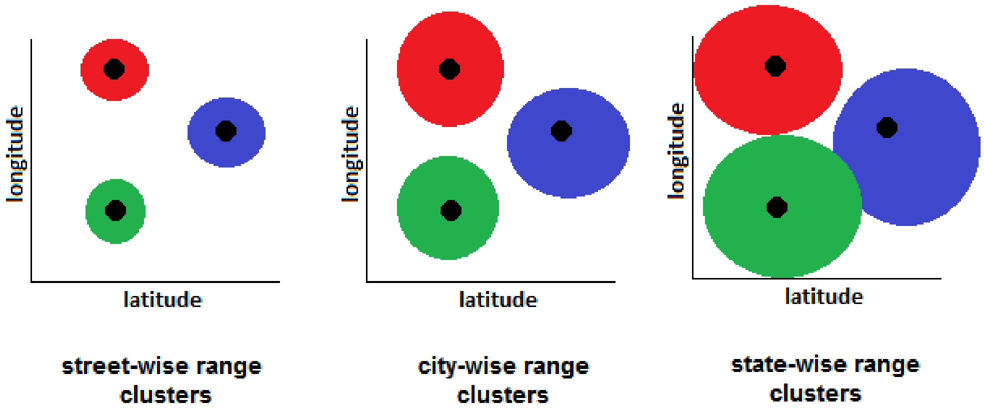
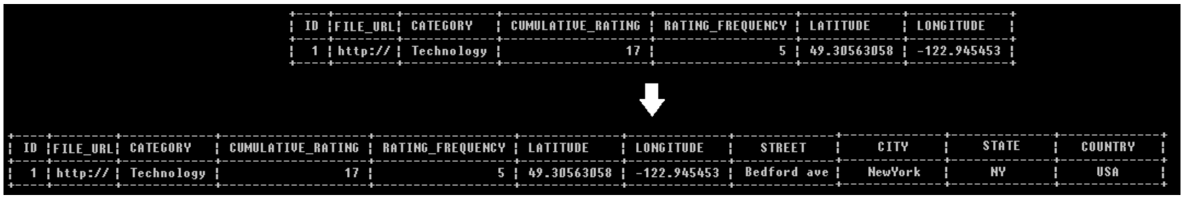


Figure 16 An example of n differently sized clusters required for each centroid to serve n different ranges of interest.

7.2 Reverse Geocoding Clustering

The term reverse-geocoding is the process of transition of a geographical location, in terms of latitude and longitude coordinates, into a human-readable address. For example, a geographical coordinates input (latitude=40.7612212, longitude=73.23211) would translate into the address of 291 Bedford Ave, New York, NY, U.S. The technology is widely used in navigation devices to provide the users readable street addresses which are easier to understand by the end users. Numerous online mapping service providers, such as Google Maps, provide a direct way to use their reverse-geocoding service via a simple HTTP request. As the multimedia content of LIVESGEO is automatically tagged with latitude and longitude coordinates at the time of creation, reverse-geocoding allows more efficient and precise ways to geographically cluster the geo-tagged contents. Our framework decomposes a reversely geocoded address into individual address components and stores these components as the attribute values of a geo-tagged multimedia content in the database, as depicted by Figure 17.



```
+----+--------+------------+-------------------+------------------+-------------+-------------+
| ID |FILE_URL| CATEGORY   | CUMULATIVE_RATING | RATING_FREQUENCY | LATITUDE    | LONGITUDE   |
+----+--------+------------+-------------------+------------------+-------------+-------------+
|  1 | http://| Technology |                17 |                5 | 49.30563058 | -122.945453 |
+----+--------+------------+-------------------+------------------+-------------+-------------+

+----+--------+------------+-------------------+------------------+-------------+-------------+-------------+-----------+------------+------------+
| ID |FILE_URL| CATEGORY   | CUMULATIVE_RATING | RATING_FREQUENCY | LATITUDE    | LONGITUDE   |   STREET    |   CITY    |   STATE    |  COUNTRY   |
+----+--------+------------+-------------------+------------------+-------------+-------------+-------------+-----------+------------+------------+
|  1 | http://| Technology |                17 |                5 | 49.30563058 | -122.945453 | Bedford ave |  NewYork  |     NY     |    USA     |
+----+--------+------------+-------------------+------------------+-------------+-------------+-------------+-----------+------------+------------+
```

Figure 17 An example of how the embedded latitude and longitude coordinates of a content is reversely geocoded and stored as the individual address component values**.**

This allows the system to filter contents into different spatial range clusters by simply selecting instances based on the values of the address component attributes. This eliminates the requirement of having n pre-clusters for n different spatial ranges. Moreover, it eliminates the ambiguity of an instance falling into more than one geographical cluster, as each instance is given an exact address. The framework can cluster contents into address-specific groups rather than ambiguous clusters that might overlap different street, city, state or country zones. Figure 18 illustrates the advantage of using reverse-geocoding over a clustering algorithm. However, reverse-geocoding does not guarantee perfect address translation, as it is only an attempt to find the closest addressable location with a certain level of tolerance. We will delve into the performance of reverse-geocoding in a later section.

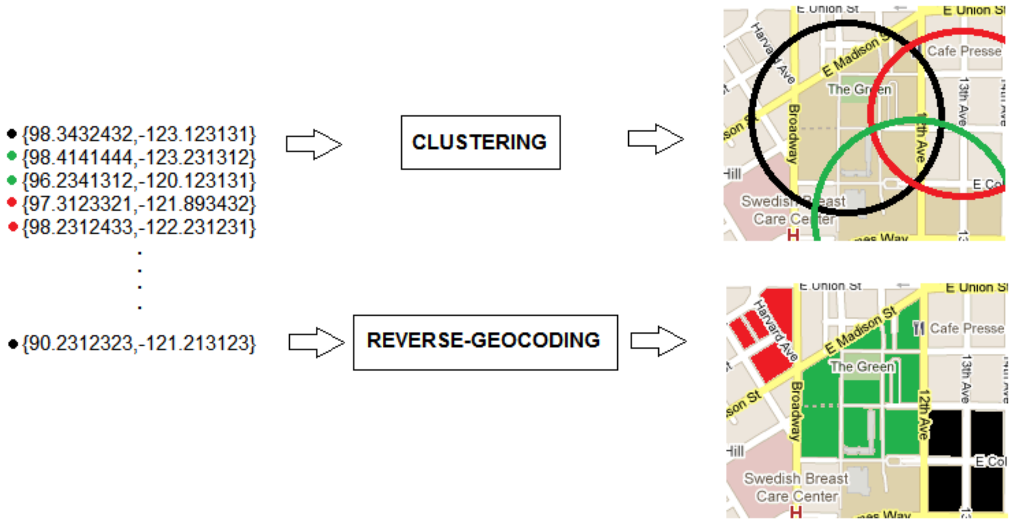


Figure 18 Clustering the geo-tagged contents using a clustering algorithm and reverse-geocoding. Notice how the reverse-geocoded instances provide more fine-grained clusters.

7.3Great Circle Distance

The term great circle means the intersection of a sphere and a plane which passes through the center point of the sphere. Simply, the great circle is the largest circle that you can get from a sphere by cutting the sphere into half through its center. Between a set of two unique points on the surface of the Earth, there always is a unique great circle that passes through these two points. These two points separate the great circle into two arcs and the shorter arc is defined as the great-circle distance [38] between the two points. As the Earth is nearly spherical, the mathematical formulae for great circle distance are commonly used in technologies related to navigation to calculate a distance between two points on the Earth.

7.3.1 Haversine and Vincenty Formulae

Let's first define the variables used in great circle distance formulae:

- Let α_s and ρ_s be latitude and longitude of a starting point on a map.
- Let α_e and ρ_e be latitude and longitude of an end point on a map.
- Let $\Delta\alpha$ and $\Delta\rho$ be their differences, respectively.
- Let $\emptyset$ be the central angle between starting and end points.
- Let $\mathbf{d}$ be the great circle distance between two points.

The central angle, $\emptyset$, can be calculated using the spherical law of cosines, Equation 7.1.

$$\emptyset = \cos^{-1}(\sin\alpha_s \cdot \sin\rho_s + \cos\alpha_e \cdot \cos\rho_e) \tag{7.1}$$

However, the general spherical law of cosines results in unstable values for small distances. If two points are relatively close to each other, the arc cosine value nears ± 1 and results in a large rounding error. Table 4 shows the limitation of the spherical law of cosines when applied to short distances. To achieve better accuracy, we looked into two alternative formulae, which are the Haversine [39] and Vincenty [40] formulae, widely used in the field of cartography.

α_s and ρ_s	α_e and ρ_e	True distance (km)	Spherical law of cosines (km)
49.25161969428559,-123.05412903428078	49.251649135854336,-123.05418267846107	0.005	3.0073026354275156
49.25195026902823,-123.0533216893673	49.25175068096708,-123.0539707839489	0.05	3.0072943142182593
49.251974779787105,-123.053447753191	49.25176818872402,-123.05476740002632	0.1	3.007296942240496

Table 4 The table shows the results of applying the spherical law of cosines on short distances. Notice the significant inaccuracy of the formula.

The Haversine formula provides an improved numerical accuracy when computing a spherical triangle whose length is very small. [41] outlines the detail process of formulating the law of Haversine from the general spherical law of cosines. Equation 7.2 defines the Haversine formula for calculating the central angle using the coordinate values.

$$\emptyset = 2 \cdot \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\Delta\alpha}{2}\right) + \cos\alpha_s \cdot \cos\alpha_e \cdot \sin^2\left(\frac{\Delta\rho}{2}\right)}\right) \tag{7.2}$$

The Vincenty formula is another popular formula for calculating the great circle distance. It differs from the Haversine formula as the formula is based on the assumption that the Earth is an oblate spheroid rather than a sphere. [40] outlines the detail process of formulating the law of Vincenty. Equation 7.3 defines the Vincenty formula for calculating the central angle using the coordinate values.

$$\emptyset = 2 \cdot \sin^{-1}\left(\frac{\sqrt{(\cos\alpha_e \sin\Delta\rho)^2 + (\cos\alpha_s \cdot \sin\alpha_e - \sin\alpha_s \cdot \cos\alpha_e \cdot \cos\Delta\rho)^2}}{\sin\alpha_s \cdot \sin\alpha_e + \cos\alpha_s \cdot \cos\alpha_e \cdot \cos\Delta\rho}\right) \tag{7.3}$$

The great circle distance is then:

$$d = \mathrm{R} \cdot \emptyset \tag{7.4}$$

where $\emptyset$ can be calculated either using Haversine or Vincenty and R is the mean radius of the Earth defined by International Union of Geodesy and Geophysics [42], which is 6378.137km.

Although the correctness of the Haversine and Vincenty formulae has been validated through numerous academic literatures [43,44], we could not find any work on a direct comparison of two formulae. Thus, we have performed a preliminary evaluation of comparing two formulae, as we need the optimal great circle distance formula to justify the use of reverse-geocoding.

7.3.2 Evaluation and Results

A field experiment was done on an observably flat ground recording the geographical coordinates of a set of endpoint pairs along with the actual distance between each endpoint pair. The great circle distances were then calculated for each pair using the Haversine and Vincenty formulae. These calculated distances were compared with the actual distance of each pair. The experimental distances of the 12 sample pairs ranged from 5m to 100m. Table 5 shows a portion of the experimental dataset.

α_s and ρ_s	α_e and ρ_e	True distance (km)	Haversine distance (km)	Vincenty distance (km)
49.25161,-123.05412	49.25164,-123.05418	0.005	0.0050926670473707	0.0050926670474977
49.25121,-123.05297	49.25122,-123.05284	0.01	0.0090180952231876	0.0090180952231281
49.251719,-123.05338	49.25169,-123.05315	0.015	0.0173095436244119	0.0173095436244143

Table 5 The table shows a portion of the experimental data. Each sample pair shows its actual distance and estimated great circle distances.

Figure 19 shows the distance estimates of the Haversine and Vincenty formulae compared against the actual distances. The difference, in the average deviation from the actual distance, between two formulae was $6.7154 \cdot 10^{-14}$km, which was insignificant. This indicates that the accuracy of the formulae is identical for practical purposes.

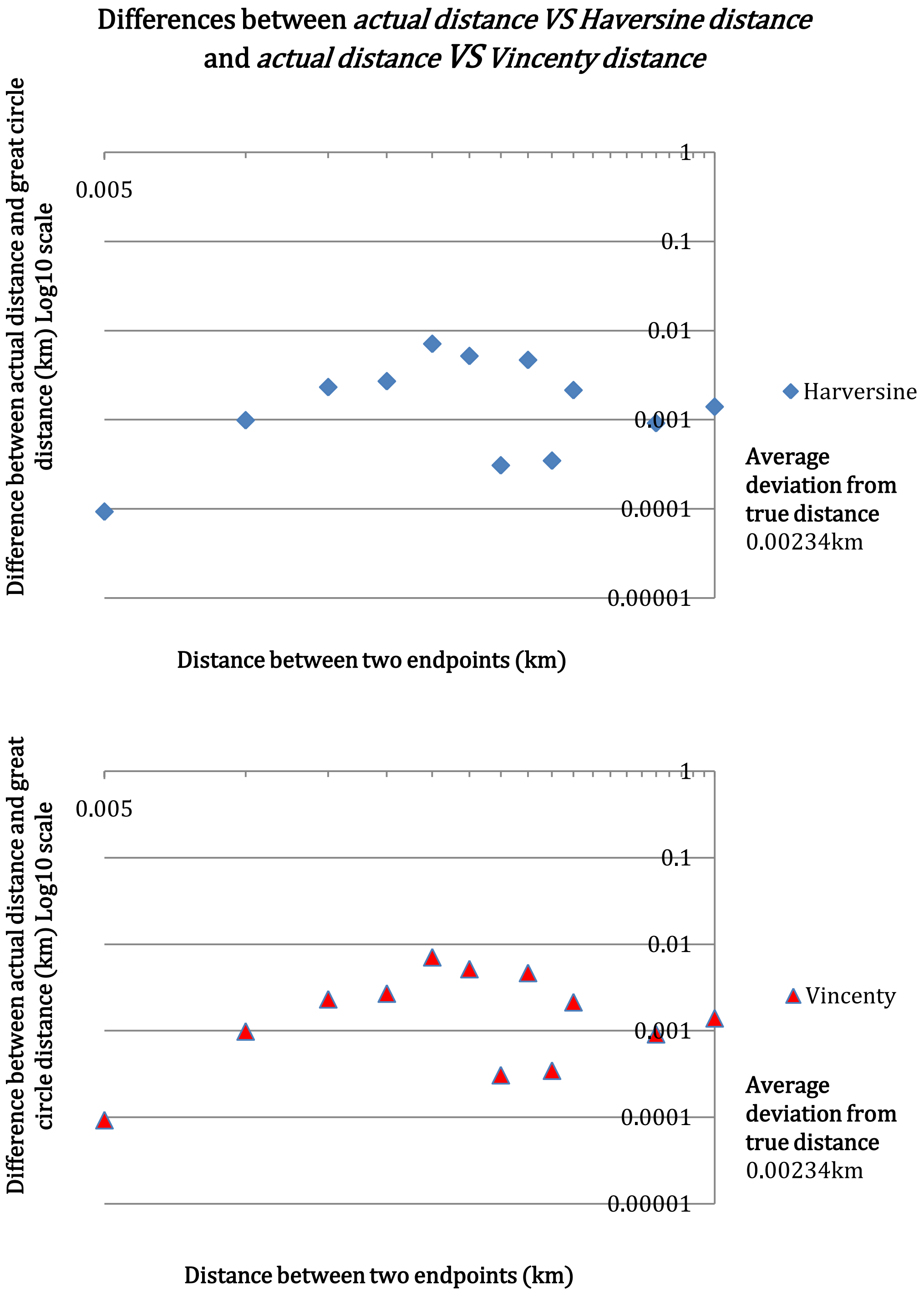


Figure 19 The differences between the actual distances and the great circle distances calculated using the Haversine and Vincenty formulae on the 12 sample pairs. Each

marker indicates an endpoint pair sample. The distance between the endpoints increases from left to right on the x-axis of the graph, 0.005km to 1km, respectively. The difference between the actual distance and the estimated great circle distance is shown in log 10 scale on the y-axis. Notice that the average deviation from the actual distance is practically the same for both Haversine and Vincenty.

7.4Evaluation

In this section we evaluate the accuracy of reverse-geocoding in order to justify the use of reverse-geocoding as the clustering engine of our framework. As reverse-geocoding is an estimate attempt to find the closest addressable location with a certain undisclosed level of tolerance, we evaluate the accuracy of reverse-geocoding on sample locations in the Greater Vancouver area.

7.4.1 Experimental Setup

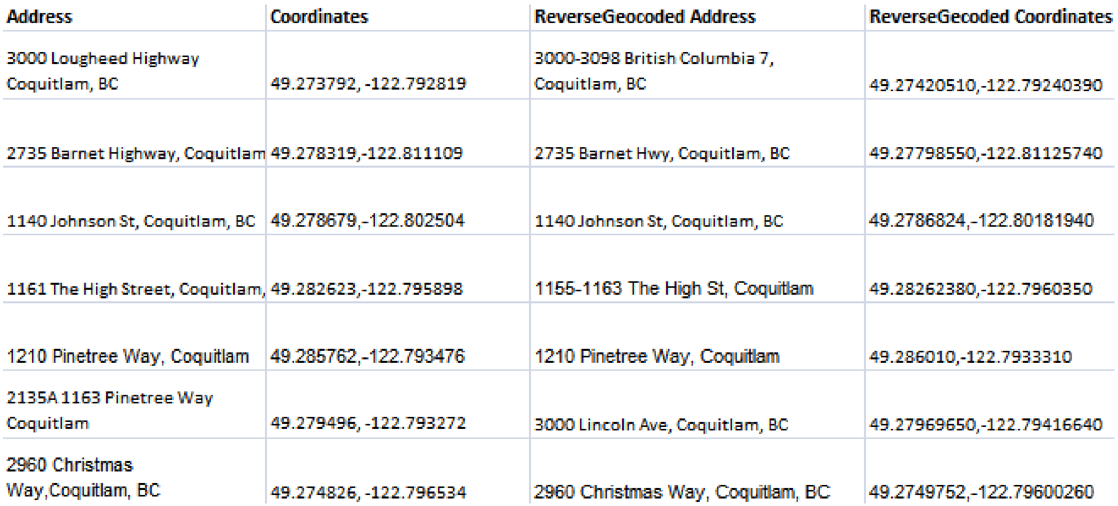
Study Area: The location sampling was conducted in the Greater Vancouver area.
Sample Data: Each sample location datum was gathered by obtaining the latitude and longitude coordinates of the actual location using a GPS device.

70 random sample locations around the cities of the Greater Vancouver area were selected including: West Vancouver, North Vancouver, Vancouver, Burnaby, New Westminster, Richmond, Port Moody, Coquitlam and Surrey. At each sample location, the latitude and longitude coordinates were recorded along with the nearest observable building's address. As reverse-geocoding returns an estimated nearest addressable location of the given geographical coordinates, we can compare the address of the true nearest building we have observed and the address of the estimated nearest location calculated by reverse-geocoding to evaluate it accuracy. Therefore, 70 random locations' latitude and longitude coordinates were fed into Google's reverse-geocoding API and the returned results of the estimated addressable locations were compared with the true nearest addressable locations we have observed for each sample.

7.4.2 Evaluation Metric

The deviation of each location sample's true nearest address from reverse-geocoding's estimated nearest address was measured in terms of the great circle distance using the Haversine formula, as discussed in 7.3 in detail. In order to do so:

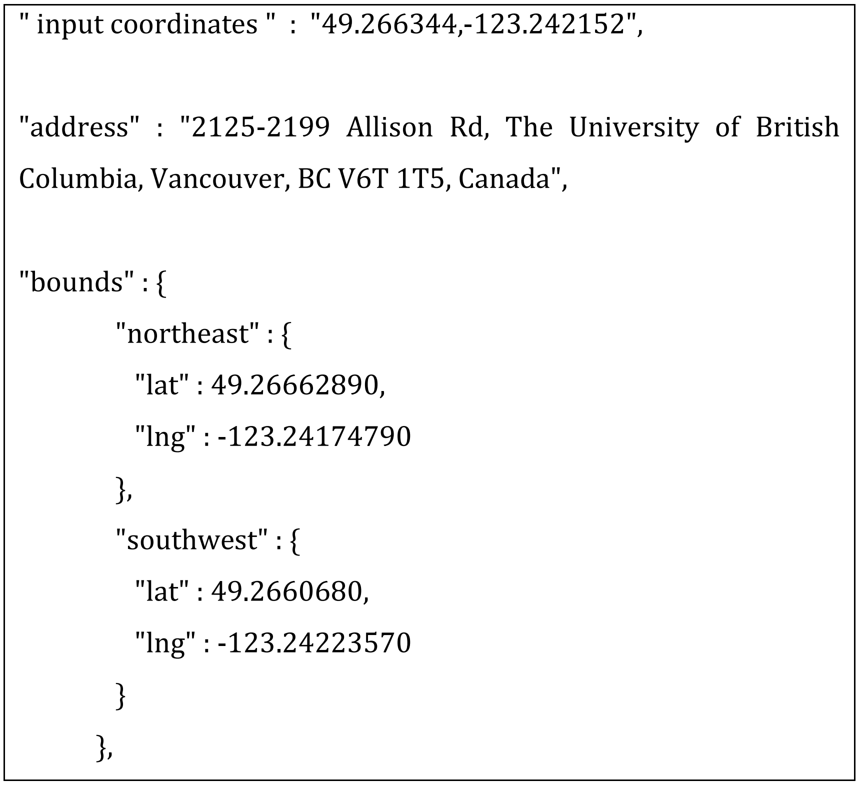
1. For each location sample, the latitude and longitude coordinates of reverse-geocoding's estimated nearest address are obtained by applying geocoding, *which translate an address to geographical coordinates*, on the estimated address. We are guaranteed to obtain the true geographical coordinates of the given address because geocoding is not an estimate, as opposed to revser-geocoding. See Figure 20 and 22.
2. The great circle distance between the latitude and longitude coordinates of the true nearest address and the reverse-geocoded addresses is calculated using the Haversine formula.



Address	Coordinates	ReverseGeocoded Address	ReverseGecoded Coordinates
3000 Lougheed Highway Coquitlam, BC	49.273792, -122.792819	3000-3098 British Columbia 7, Coquitlam, BC	49.27420510,-122.79240390
2735 Barnet Highway, Coquitlam	49.278319,-122.811109	2735 Barnet Hwy, Coquitlam, BC	49.27798550,-122.81125740
1140 Johnson St, Coquitlam, BC	49.278679,-122.802504	1140 Johnson St, Coquitlam, BC	49.2786824,-122.80181940
1161 The High Street, Coquitlam,	49.282623,-122.795898	1155-1163 The High St, Coquitlam	49.28262380,-122.7960350
1210 Pinetree Way, Coquitlam	49.285762,-122.793476	1210 Pinetree Way, Coquitlam	49.286010,-122.7933310
2135A 1163 Pinetree Way Coquitlam	49.279496, -122.793272	3000 Lincoln Ave, Coquitlam, BC	49.27969650,-122.79416640
2960 Christmas Way,Coquitlam, BC	49.274826, -122.796534	2960 Christmas Way, Coquitlam, BC	49.2749752,-122.79600260

Figure 20 The experimental dataset of the sample locations. Each sample is composed of the true nearest address, the actual coordinates, the estimated nearest address from reverse-geocoding, and the geographical coordinates of the reverse-geocoded address.

When reverse-geocoding cannot convert given latitude and longitude coordinates to an exact estimated addressable location, it returns an address range that a given location is estimated to be located within it. For example, reverse-geocoding returned an address range of ‘2125-2199 Allison Rd, Vancouver, BC’ for the input coordinates (latitude=49.266344, -123.242152), while its true nearest address was ‘2155 Allison Road, Vancouver, BC’. Google reverse-geocoding API also returns the bounding coordinate box, in terms of latitude and longitude coordinates, which will fully contain the corresponding address range. Figure 21 shows an example of the address range and the bounding coordinate box returned for the previously mentioned coordinates. For such a case, we have used the center of the bounding coordinate box as the geographical coordinates of the reverse-geocoded address.



```
" input coordinates " : "49.266344,-123.242152",

"address" : "2125-2199 Allison Rd, The University of British
Columbia, Vancouver, BC V6T 1T5, Canada",

"bounds" : {
        "northeast" : {
          "lat" : 49.26662890,
          "lng" : -123.24174790
        },
        "southwest" : {
          "lat" : 49.2660680,
          "lng" : -123.24223570
        }
      },
```

Figure 21 The result returned by Google’s reverse-geocoding API when the given geographical coordinates cannot be translated into an exact estimated nearest addressable location. Notice that it returns the bounding coordinate box that encompasses the estimated address range.

Figure 22 depicts the process of evaluating the deviation between the true address and reverse-geocoding's estimated address of each location sample.

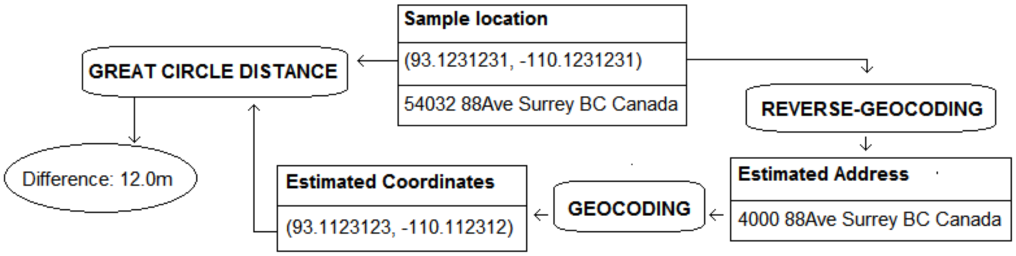


Figure 22 Evaluating the accuracy of reverse-geocoding.

7.4.3 Results

In order to investigate the performance of reverse-geocoding in terms of its prediction accuracy: first, we look at if reverse-geocoding can identify the nearest addressable location of a given sample, second, whether it correctly or incorrectly identifies it, and finally, the standard deviation of the incorrectly identified samples. Figure 23 shows the proportion of the correctly identified addresses and the incorrectly identified addresses.

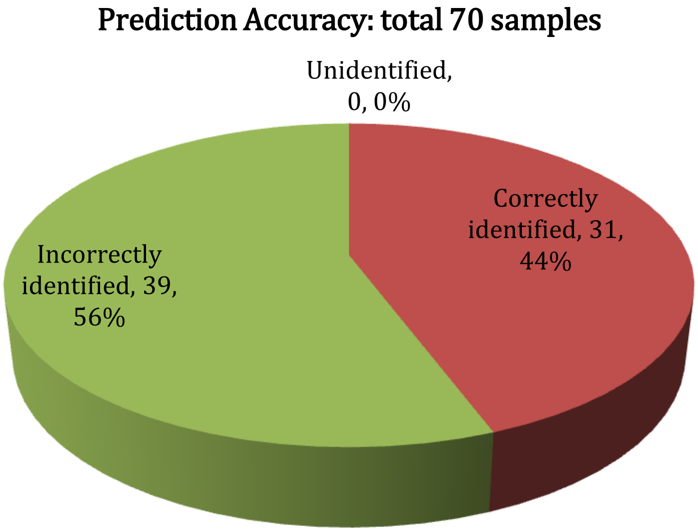


Figure 23 The proportion of the correctly and incorrectly identified addresses.

Although there was not a single case where reverse-geocoding failed to identify given geographical coordinates, more than half of total 70 samples were incorrectly identified, specifically 56 percent. This result confirms that reverse-geocoding is significantly error prone in terms of predicting the true nearest addressable location from given geographical coordinates. We noticed the trend that most of the correctly identified locations were near commercial buildings and the incorrectly identified locations were near private and residential buildings.

We also looked at the distance deviations of the 39 incorrectly identified addresses from the true nearest addressable locations in order to draw a more insightful conclusion about the accuracy of reverse-geocoding. The distances were measured using the Haversine formula. Figure 24 shows these deviations in terms of the great circle distance.

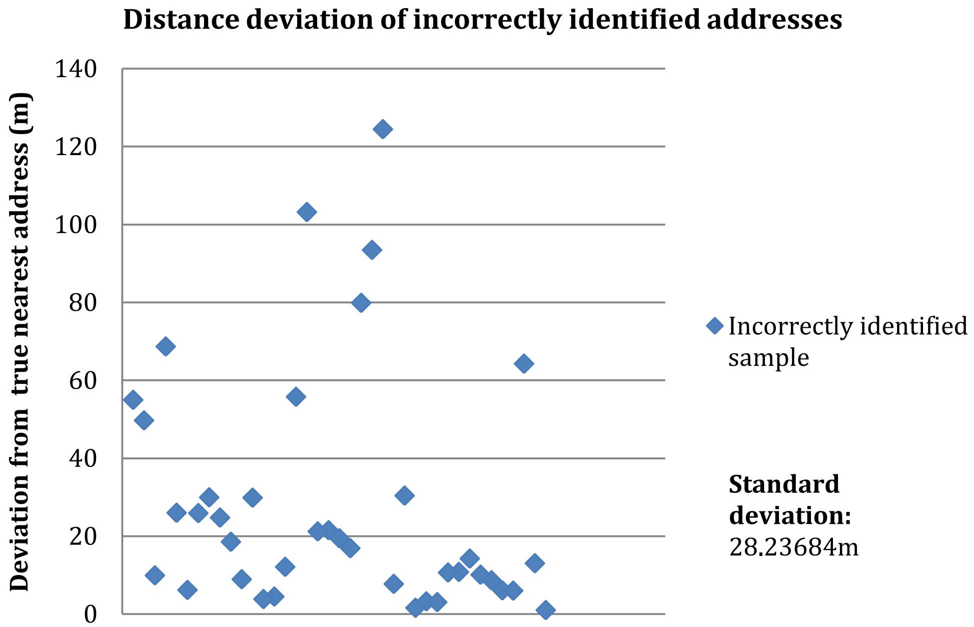


Figure 24 The distance deviation of the incorrectly identified address from the true nearest address for each incorrectly identified location sample.

We noticed that the standard deviation of the incorrectly identified addresses was not significant. Only about 9 locations deviated more than 40m from the true nearest addresses and 30 locations deviated less than 40m. On average, the deviation was 28.24m, which is a distance that would not significantly alter the geographical clustering performance of our framework. The results show how reverse-geocoding is accurate to the nearest addressable location with an acceptable error rate and is indeed an effective way to cluster our geo-tagged content spatially.

7.5Applications

We provide examples of applications that utilize the spatial content cluster framework, which offers a way to access the contents in LIVESGEO as geographical groups, to provide a glimpse of what kinds of interesting applications can be designed leveraging this framework.

7.5.1 Geographical Coverage Specific Popularity Analysis

In this section, we briefly describe an application that can be built upon the spatial content cluster framework to provide geographical coverage specific popularity of contents based on the user's current location. Geographical coverage specific popularity of a content will represent the content's reputation and quality among other contents that are in vicinity to the user's current location within different geographical coverage rather than a simple global popularity. Figure 25 shows an example of the geographical coverage specific popularity.

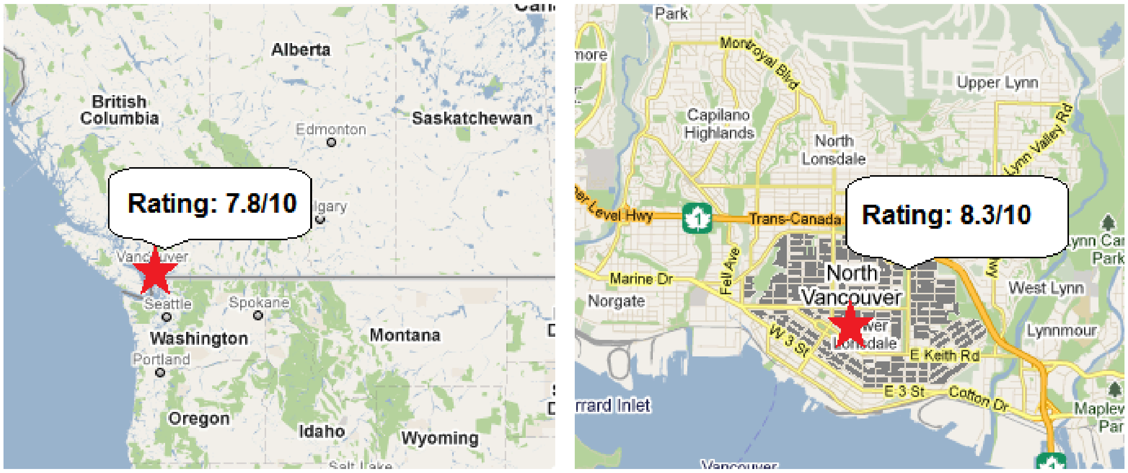


Figure 25 The national-level popularity and the city-level popularity of the same content. Notice that its relative popularity will vary depending on the size of the cluster, or the geographical coverage, it belongs to.

The spatial content cluster framework can return a set of geographically clustered contents that is clustered based on the spatial specific parameters passed by a calling application. For example, an application can request the framework to return a set of contents that has BC as the value of the province attribute and Vancouver as the value of the city attribute. As each content will have its rating information, our popularity measurement can be applied on a given set to measure the popularity of the contents in given geographical coverage.

7.5.2 Geographical Coverage Specific Content Layout

When the LIVESGEO mobile application displays the geo-tagged content layered on top of a virtual map, its performance can easily be bogged down if all the contents in the system are layered out at once. This can easily be circumvented with the help of the spatial content cluster framework using the same idea as the application described above. The client can initially layout a small set of contents that are in close vicinity of the user's current location. And upon the user's request, the client can layout a larger set of contents increasing the geographical coverage. For example, when a user launches a map to navigate through the contents, the application can start with the layout of the street-level cluster and increase to the city-level cluster upon the user's request.

8 Conclusions and Future Work

New ideas, policies and technologies are expeditiously introduced to amend emerging problems and improve the quality of life, and we are under demands to comprehend and learn from them to grow with the rest of the world. As learning is acquisition of knowledge or skills through utilizing given information, it is critical that one needs to be literate in order to learn effectively. However, conventional learning materials are mostly targeted for literate audiences and delivered by a means of written materials in classical educational settings making it harder for illiterate audiences to engage in learning. This thesis was motivated to help such less privileged populations and facilitate open learning environments.

We introduced a novel location-based multimedia knowledge sharing system that takes the advantages of the advanced capabilities of smart-phones and mobile data networks to allow peer driven learning. We provided high level design specifications and implementation decisions to enable the system to work most effectively and efficiently as a distributed mobile application. We provided careful analyses on the limitations and advantages of developing a learning system on mobile devices and proposed a hybrid of central server and peer to peer models to circumvent the limitations and leverage the advantages. The implementation described in this thesis represents an important initial step in exploring the use of a mobile platform and its mature ecosystem in peer driven open learning.

LIVESGEO enables the users to rate contents as it adapts to the idea of enforcing community introspection by rewarding the users providing higher quality contents. Thus, we provided our own novel mathematical framework to characterize the true popularity of a rated content relative to other rated contents. We evaluated its performance by performing empirical comparisons against popular rating frameworks. We observed that our popularity framework successfully characterised the underlying reputation and quality of contents.

We also explored the applicability and usefulness of the geo-tagged content accumulated by LIVESGEO. To demonstrate this, we abstract the system into the spatial content cluster framework to offer a way to access the contents in LIVESGEO as geographical groups. We have used the reverse-geocoding technology over classical clustering algorithms to geographically cluster contents with providing the explanations and rationale behind the design of our clustering framework. In order to justify the use of reverse-geocoding to cluster contents geographically, we provided empirical analyses of reverse-geocoding by making novel use of the Haversine and Vincenty formulae. First, the performance of the Haversine and Vincenty formulae were evaluated. The results indicated the accuracy of both formulae is identical for practical purposes. Second, the address translation accuracy of reverse-geocoding was evaluated by using the aforementioned formulae. The results indicated that reverse-geocoding was highly reliable with an acceptable error rate. We also provided examples of interesting applications which utilize the power of our content clustering framework.

The work presented here may be continued in myriad possible directions. We have not published the application on mobile market places yet. More detail and thorough analyses and evaluations of the system can be done after the application gains a reasonable sized user base. For instance, with a practical number of users, we can stress test the performance of video and audio streaming. It would also be possible to research the social aspect of peer driven open learning by performing empirical analyses of the real world learning trends and learning content popularity. Moreover, LIVESGEO, which is a preliminary implementation, can undergo further improvements.

We used two popular online movie rating frameworks to evaluate the performance of our popularity framework. It may be interesting to investigate how it would perform against a number of different rating frameworks. Although, the performance was satisfying, more extensive evaluations may reveal negative performances. The popularity measurement can take other factors into account, such as the number of times a content was shared through popular social networks, such as Facebook and Twitter.

We selected reverse-geocoding as the engine of our spatial content cluster framework. The framework may provide the geographical clusters of contents more efficiently, if the servers themselves are clustered geographically. For instance, the contents of Vancouver can physically be stored separately from the contents of Seattle. This would allow a single request to return the geographically clustered contents of a particular city without having to search all content instances in the agglomerated dataset of a central server and with minimal network latency.

Bibliography

[1] The United Nations Educational, Scientific and Cultural Organization, "The Plurality of Literacy and its Implications for Policies and Programmes," UNESCO, 2004.

[2] S. T. Vuong, et al., "LIVES: Learning through Interactive Voice Educational System," *Proceedings of the IADIS International Conference on Mobile Learning*, vol. 19, Mar. 2010.

[3] S. T. Vuong, J. Schroeder, M. S. Alam, and D. Chan, "Mobile Learning for Farmers via LIVES," in *In Sixth PAM-Commonwealth Forum Open Learning*, 2010.

[4] S. Kuljeet, "LIVESMOBILE: Extending the LIVES system to support SMS and MMS for mobile learning," Master's thesis, The University of British Columbia, Vancouver, 2010.

[5] T-Mobile. (2011, Jul.) T-Mobile Knowledge Base Entry. [Online]. http://developer.t-mobile.com/loadKbaseEntry.do?solutionId=2001

[6] J. Deeq and D. George, "Literacy: State of the Nation," National Literacy Trust, Research Report, 2010.

[7] D. George and C. Christina, "Literacy Changes Lives: An advocacy resource," National Literacy Trust, London, Research Report, 2008.

[8] Central Intelligence Agency. (2022, Jul.) GDP - per capita (PPP) vs. Literacy. [Online]. http://www.indexmundi.com/g/correlation.aspx?v1=67&v2=39&y=2003

[9] P. Cottrell. (2010, Jul.) The Daily Observer: Understanding the importance of literacy. [Online]. http://www.thedailyobserver.ca/ArticleDisplay.aspx?e=2931215

[10] F. Ross and M. Ronald, "The Importance of Functional Literacy: Reading and Math Skills and Labour Market Outcomes of High School Drop-outs," Government of Canada Research Paper, 2006.

[11] R. N. Marcus, "The Silent Epidemic: The Health Effects of Illiteracy," *The New English Journal of Medicine*, vol. 355, pp. 339-341, Jul. 2006.

[12] The United Nations Educational, Scientific and Cultural Organization, "Education for All: Literacy for Life," UNESCO, Paris, Global Monitoring Report, 2005.

[13] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2010-2015," Cisco White Paper, 2011.

[14] eMarketer Digital Intelligence. (2011, Jul.) Canadian Mobile Subscription to Climb 20% by 2014. [Online]. http://www.emarketer.com/Article.aspx?R=1007747

[15] mobiThinking. (2011, Jul.) Global mobile statistics 2011: all quality mobile marketing research, mobile Web stats, subscribers, ad revenue, usage, trends…. [Online]. http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats

[16] R. Meier, *Professional Android 2 Application Development*, 2nd ed. Indianapolis, U.S: Wiley Publishing, Inc, 2010.

[17] A. K. Saha, "A Developer's First Look at Android," *Linux for You*, pp. 48-50, Jan. 2008.

[18] Google. (2011, Jul.) Android Developers. [Online]. http://developer.android.com/index.html

[19] Wireless Federation. (2011, Jul.) Global Android market share increases to 35% in Q1. [Online]. http://wirelessfederation.com/news/74356-global-android-market-share-increases-to-35-in-q1/

[20] N. Gandhewar and R. Sheikh, "Google Android: An emerging software platform for mobile devices," *International Jounral on Computer Science and Engineering*, no. Special Issue, pp. 12-17, Feb. 2011.

[21] Rysavy Research, "EDGE, HSPA and LTE: The mobile broadband advantage," 3G americas Research Paper, 2007.

[22] W. L. Tan, F. Lam, and W. C. Lau, "An Empirical Study on the Capacity and Performance of 3G Networks," *IEEE Transactions on Mobile Computing*, vol. 7, pp. 737-750, Jun. 2008.

[23] S. Parkvall, et al., "LTE-Advanced: Evolving LTE towards IMT-Advanced," in *IEEE Vehicular Technology Conference*, Calgary, 2008, pp. 21-24.

[24] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, United States of America: Morgan Kaufmann, 2005.

[25] Alexa. The Web Information Company. (2011, Jul.) www.alexa.com. [Online]. http://www.alexa.com

[26] TIME. (2011, Jul.) Look Who's Using Wikipedia. [Online]. http://www.time.com/time/business/article/0,8599,1595184,00.html

[27] REUTERS. (2011 , Jul.) Wikipedia remains go-to site for online news. [Online]. http://www.reuters.com/article/2007/07/08/us-media-wikipedia-idUSN0819429120070708

[28] J. Giles, "Internet encyclopedia go head to head," *Nature*, vol. 438, pp. 900-901, Dec. 2005.

[29] F. B. Viegas, M. Wattenberg, and K. Dave, "Studying Cooperation and Conflict between Authors

with history flow Visualization ," *Proceedings of the ACM Conference on Human Factors in Computing Systems*, vol. 6, pp. 575-582, Apr. 2004.

[30] R. Priedhorsky, et al., "Creating, Destroying, and Restroing Value in Wikipedia," in *ACM Group 2007 Conference Proceedings*, New York, 2007.

[31] DigitalGreen. (2011, Jul.) Digital Green. [Online]. http://www.digitalgreen.org

[32] R. Gandhi, R. Veeraraghanvan, K. Toyama, and V. Ramprasad, "Digital Green: Participatory Video for Agricultural Extension," *International Conference on Information and Communcation Technologies and Development*, pp. 1-10, May 2009.

[33] R. Gandhi, R. Veeraraghanvan, K. Toyama, and V. Ramprasad, "Digital Green: A participatory Digital Framework to Deliver Targeted Agricultural Information to Small and Marginal Farmers," in *The ASA-CSSA-SSSA International Annual Meetings*, New Orleans, 2007.

[34] S. Shah and A. Joshi, "COCO: A Web-Based Tracking Architecture for Challenged Network Environments," in *ACM DEV*, London, 2010.

[35] UpSide Learning. (2011, Jul.) UpSide Learning. [Online]. http://www.upsidelearning.com/

[36] A. Kadle, "Do You Need Games In Your Elearning Mix?: A Whitepaper by Upside Learning Solutions," UpSide Learning White Paper, 2009.

[37] J. V. Meggelen, J. Smith, and L. Madsen, *Asterisk: The Future of Telephony*, 2nd ed., M. Loukides, Ed. California, United States of America: O'Reilly Media, 2005.

[38] E. W. Weisstein. (2011, Jul.) MathWorld- A Wolfram Web Resource: Great Circle . [Online]. http://mathworld.wolfram.com/GreatCircle.html

[39] R. W. Sinnott, "Virtues of the Haversine," *Sky and Telescope*, vol. 68, no. 2, p. 158, 1984.

[40] T. Vincenty, "Direct and Inverse Solutions of Geodesics on the ellipsoid with Application of Nested Equations," *Survey Review*, vol. 22, pp. 88-93, Apr. 1975.

[41] R. D. Miller, "Computing the Area of a Spherical Polygon," in *Graphics Gems IV*, P. S. Heckbert, Ed. Boston, U.S: Academic Press, 1994, ch. 2, pp. 132-133.

[42] (2011, Jul.) International Union of Geodesy and Geophysics. [Online]. http://www.iugg.org/

[43] C. M. Thomas and W. E. Featherstone, "Validation of Vincenty's Formulas for the Geodesic Using a New Fourth-Order Extension of Kivioja's Formula," *Journal of Surveying Engineering*, vol. 131, pp. 20-26, Feb. 2005.

[44] Hijmans. (2011, Jul.) Introduction to the geosphere package. [Online]. http://cran.r-

project.org/web/packages/geosphere/vignettes/geosphere.pdf