Fall 11-1-2016

An Analytics-based Framework for Engaging Prospective Students in Higher Education Marketing

Santhosh Kumar Lakkaraju
Dakota State University

AN ANALYTICS-BASED FRAMEWORK FOR ENGAGING PROSPECTIVE STUDENTS IN HIGHER EDUCATION MARKETING

A dissertation submitted to Dakota State University in partial fulfillment of the requirements for the degree of

Doctor of Science

in

Information Systems

November, 2016

By
Santhosh Kumar Lakkaraju

Dissertation Committee:
Dr. Deb Tech (Chair)
Dr. Shuyuan Deng
Dr. Ronghua Shan
Dr. Mark Hawkes

DISSERTATION APPROVAL FORM

This dissertation is approved as a credible and independent investigation by a candidate for the Doctor of Science in Information Systems degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this dissertation does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department or university.

Student Name: Santhosh Kumar Lakkaraju

Dissertation Title: **An Analytics-Based Framework for Engaging Prospective Students in Higher Education Marketing**

Dissertation Chair/Co-Chair: ______________________ Date: 11/6/16
(Dr. Deb Tech)

Committee member: ______________________ Date: 11/3/2016
(Dr. Shuyuan Deng)

Committee member: Ronghua Shan Date: 11/4/2016
(Dr. Ronghua Shan)

Committee member: Mark Hawkes Date: 11-7-16

(Dr. Mark Hawkes)

DSU Institutional Review Board

Expedited Project Approval

To: Santhosh Lakkaraju
Date: March 16, 2016

Project Title: Analytics-Based Framework for Profiling Prospective Students and Improving Prospect Engagement through Mass-Customized Communication Sequences in Higher Education

---

Approval #: 2015-2016-110

The IRB approved your project using expedited procedures as described in 45 CFR 46.110. The activity was deemed to be no greater than minimal risk, and the following expedited category from 63 FR 60364-60367 was found to be applicable to your activity:
(7) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.
One-year approval of your project will be dated starting March 16, 2016. If you require additional time to complete your project or wish to extend the activity, please submit a request for extension before February 16, 2017. The request can be submitted by email to IRB@dsu.edu. If there are any unanticipated problems involving risks to subjects or others, or if there are changes in the procedures during the study, please contact the Sponsored Programs Office at IRB@dsu.edu . Any protocol changes must be approved by the IRB prior to implementation. At the end of the project please inform the committee that your project is complete.

If I can be of further assistance, don't hesitate to let me know.

Sincerely,

Risë Smith
Chair, DSU Institutional Review Board

Dakota State University

Institutional Data Non-Disclosure Agreement

As a condition of my participation in any DSU activities or services involving DSU institutional data, I acknowledge that all institutional data, analysis results, reports, and publications (together hereinafter referred to as "research artifacts") remain the property of Dakota State University. This includes any analysis results, reports, and publications which, individually or jointly with others, I shall conduct or generate using DSU institutional data during the course of such employment or participation. I hereby agree that:

I shall maintain the confidentiality of all non-public information or research artifacts disclosed/provided to or generated by me.

I shall disclose, in writing and in responsible detail, descriptions of all research artifacts from DSU non-public information or data to the Director of the Office of Institutional Effectiveness and the Dean of Graduate Studies and Research of DSU.

I shall disclose or publish the research artifacts only with the prior written approval of the Director of the Office of Institutional Effectiveness and the Dean of Graduate Studies and Research of DSU. While each research artifact shall be deemed to be the property of DSU, DSU will consider the transfer of copyright requests for those artifacts deemed publishable.

Upon completion of my research project, I shall provide copies of research artifacts, including non-public research data, results, and materials which have been disclosed/provided to or generated by me during the project to the individuals listed in Item 3 above.

Upon completion of my employment and/or training, research artifacts including non-public research data, results and materials which have been disclosed/provided to or generated by me during the course of my employment and/or training shall remain the property of DSU and all copies of the above-listed information shall be returned to DSU. Further, I agree that I will not publish the data and/or results of my research without the explicit permission of DSU as noted above in Item 3.

santhoshlakkaraju santhoshlakkaraju 03/16/2016

Faculty, Staff or Student Signature Date

DSU Institutional Review Board

Expedited Project Approval

To: Santhosh Lakkaraju

Date: August 5, 2016

Project Title: Analytics-Based Framework for Profiling Prospective Students and Improving Prospect Engagement

---

Approval #: 2016-2017-101

The IRB approved your project using expedited procedures as described in 45 CFR 46.110. The activity was deemed to be no greater than minimal risk, and the following expedited category from 63 FR 60364-60367 was found to be applicable to your activity:
(7) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.
One-year approval of your project will be dated starting August 5, 2016. If you require additional time to complete your project or wish to extend the activity, please submit a request for extension before July 5, 2017. The request can be submitted by email to IRB@dsu.edu. If there are any unanticipated problems involving risks to subjects or others, or if there are changes in the procedures during the study, please contact the Sponsored Programs Office at IRB@dsu.edu . Any protocol changes must be approved by the IRB prior to implementation. At the end of the project please inform the committee that your project is complete.

IRB approval #2016-2017-101 is applicable only to data collected after receipt of this letter. Survey data collected prior to IRB approval is out of compliance with Office of Human Rights Protection (OHRP) 45 CFR 26 as well as DSU Policy 04-03-00 and cannot be used for research purposes or otherwise disseminated.

If I can be of further assistance, don't hesitate to let me know.

Sincerely,

Jack H. Walters, Ph.D.
Chair, DSU Institutional Review Board

DEDICATED TO MY PARENTS.

ACKNOWLEDGMENT

I would like to express sincere gratitude to my supervisors Dr. Deb Tech (Chair), Dr. Omar El-Gayar and Dr. Surendra Sarnikar (advisors), Dr. Shuyuan Deng, Dr. Mark Hawkes and Dr. Ronghua Shan (Committee Members) and Dr. Amit Deokar (Associate Professor) for their patience and never ending support. I thank my dissertation chair Dr. Tech, for spending time and sharing her knowledge with me. Dr. Omar El-Gayar played a significant role in my student life and corrected my way of approaching research. I would like to thank Dr. Ronghua Shan and Dr. Mark Hawkes for being available and extending their support.

I would also like to acknowledge the following for their assistance and motivation in helping me successfully complete all the requirements for the D.Sc.IS program. Firstly, I would like to thank my Parents: Lakkaraju Venkata Ramana Murthy and Lakkaraju Lakshmi Parvathi, my grandparents Lakkaraju Venkata Subbarao and Sarada Devi, my wife Sravanthi for her continuous support, my brother Sandeep Lakkaraju and my sister-in-law Monica for their never-ending love and support on all the things that I wish to do with focus. I felt relieved and motivated from the adorable smiles I get from my 7-month old niece Sloka. Without their love and patience, I wouldn't be able to accomplish this. I would also like to extend my thanks to all my family members and friends for their encouragement and assistance throughout my academic life.

I will treasure all the precious moments I have shared with the faculty, staff, and friends especially my family at Dakota State University.

ABSTRACT

Analytical models are already in wide use in the e-commerce marketing sector, however, their use in higher education marketing is limited. Many higher education institutions are struggling to establish a reliable and sustaining engagement strategy with their prospects' due to the lack of awareness about the kind of information a prospective student might be interested in receiving during the decision-making period. Before purchasing a higher education program, prospects invest a significant amount of time researching different programs by visiting the institutional websites. The use of information generated in such process by the institution is generally limited to reporting and website optimization purposes.

To efficiently engage the prospective students, there is a need for analytical models that would extract the prospect navigational behavior on the website and help the institutions analyze the prospects' needs. Although institutions are adopting newer engagement strategies, delivering the right information that addresses individual prospect's needs remains a difficult problem.

The objective of this research is to design and test an analytics framework that would significantly aid higher education institutions in identifying prospective students' interests and optimize their engagement strategies. This research intends to evaluate the framework by extracting the traces of prospective student information hidden in institutional server logs, profile the prospects based on the pre-inquiry, inquiry and post inquiry phases and further discuss the impact of such prospective student profiles on the institutional efforts in engaging the prospects and overall institutional advancement. The framework will help the institutions to make use of web content mining to extract the contextual profiles from the institutional web pages and apply the web usage mining to extract the prospect navigational profiles from web server logs and compare them to suggest the mass-customized communication sequences.

Finally, the framework is tested in a public university and contributed to the institutional marketing efforts to better engage the prospects in the higher-education sales cycle.

Declaration

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

____

Santhosh Lakkaraju

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

This chapter provides an overview of the existing prospect engagement problem in higher education institutions and a solution to address the problem.

1.1 Background

Prospective student acquisition is a prominent issue in higher education marketing. Noel-Levitz (2012) estimated that higher education institutions may lose as much as 75% of the prospects from the inquiry phase to the application phase. Another study reported that 80% of the students who decide to apply to a program were influenced by the post-inquiry communications they had received from the higher education institutions (Aarinen, 2012). It is evident from earlier studies that the higher education institutions tend to receive a larger number of inquiries from prospects' but there is a steady decrease in the inquiry to application ratio (Hemsley-Brown & Oplatka, 2006; Moogan, 2011; Morris, 2009).

The purchase, enrollment, is the only visible part of a more complex decision-making process from the consumer. In general, the consumer decision-making process consists of five different phases that drive consumers throughout their purchase process. Kotler & Armstrong (2006) defined the five consumer buying decision-making phases as Need Recognition, Information Search, Evaluation of Alternatives, Making a Decision and Post Purchase Behavior. Higher education institutions tackle the first four consumer buying decision-making phases through the four phases of the admission funnel. The admission funnel primarily consists of the awareness, inquiry, and application and admissions phases.

The awareness phase involves different marketing techniques the institutions rely on to reach out to prospects. The awareness phase addresses the need recognition and information search phases by providing relevant information on institutional websites.

During the inquiry phase, a prospect tends to look for potential information on the institutional website and makes an inquiry by filling out the inquiry form. The institution responds to those inquiries by sending out different communications to prospects. These communications play a critical role in helping prospects to make a decision to apply or not. The inquiry phase primarily targets the evaluation of alternatives and decision-making phases.

In the application phase, the institutions receive an application from the prospect. The cycle ends with the admission phase where the prospect would receive a decision on the application from the institution.

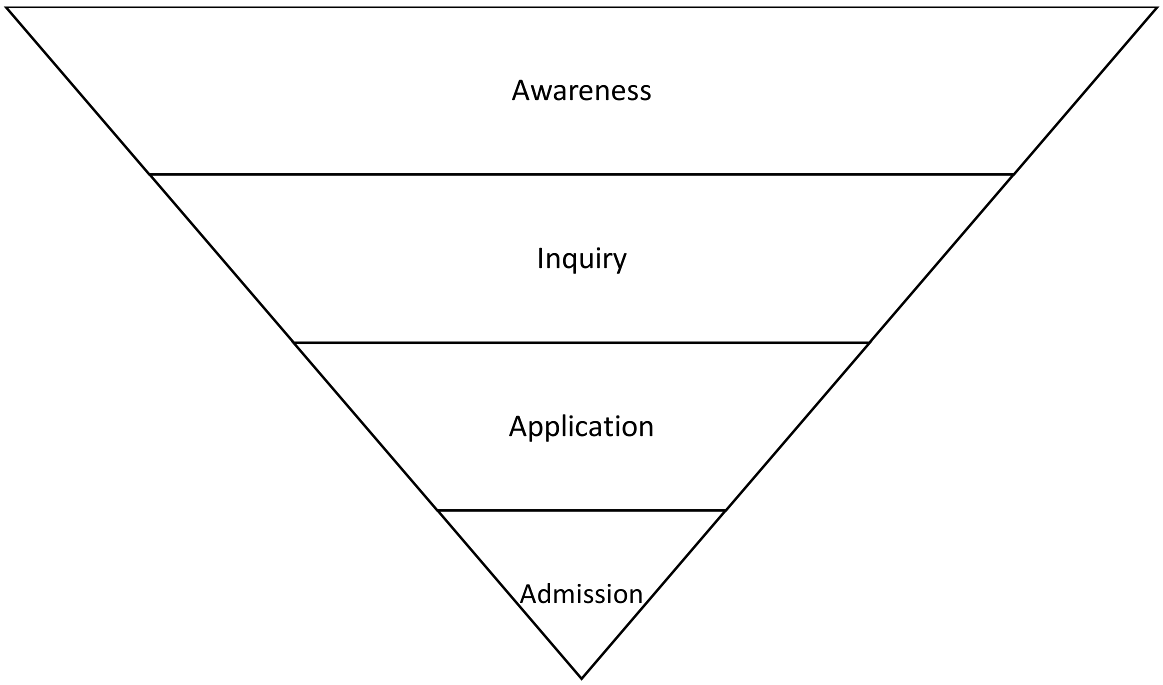

Figure 1: Admission Funnel

Institutions need to have a local and global footprint as well as maintain their brand value to attract prospects. To better market themselves and increase awareness in their targeted student population, higher education institutions are employing different techniques like online (or) pay-per-click marketing and print media to increase their local and global awareness. These marketing efforts drive prospects to the institution’s website where the prospect searches for additional information and may make an inquiry. An inquiry from a prospect involves a request to know more information about a program and can be interpreted as interest in the program. Because of the exponentially growing educational market and varying prospective student behavior, institutions receive a larger number of inquiries from prospects about a specific program than the actual enrollments into that program (Hemsley-Brown & Oplatka, 2006; Moogan, 2011; Morris, 2009).

Moogan (2011) articulated that due to the lack of awareness about the kind of information a prospective student might be interested in receiving during the decision-making period, many educational institutions are losing potential prospects. Existing models in the higher education industry tend to rely on the post-inquiry information to make their assessments and predictions.

Earlier studies investigated several key decision-making variables of a prospective student (Aarinen, 2012; Moogan, 2011; Moogan, Baron, & Harris, 1999; Schäfer & Kummer, 2013), and some studies investigated the current student demographics to predict prospective student enrollment (Desjardins, 2002; Goenner & Pauls, 2006; Tareef & Balas, 2009). Other research studies developed predictive models using prospective student geo-demographic information collected through the online inquiries and estimated the prospective student enrollment rates (Goenner & Pauls, 2006; Michael, 1990; Morris, 2009). However, most of the online inquiries that educational institutions receive are incomplete, capturing limited data, which results in inaccurate or biased predictions (Dupaul, 2010).

The timing of subsequent email communications to the prospect may be determined by the anticipated entry date information from the inquiry. A communication system structured around such assessments and predictions lack the ability to address the prospect's need.

One way to determine the prospect's interest is to extract the pre-inquiry and post-inquiry navigational behavior from the server logs. In the process of making an inquiry, the prospect leaves trails of critical navigational information that is stored in the server logs. The navigational information relates to the information accessed by the prospect from the institutional website before making an inquiry and the information accessed after making an inquiry. The information accessed before making an inquiry is considered as pre-inquiry navigational behavior and the information accessed after making an inquiry is considered as post-inquiry navigational behavior. The institutions capture this information and store them in the form of server logs. This information can be used to construct prospect profiles.

This research focuses on addressing the inquiry to application gap by targeting the prospect engagement. Instead of relying on the post-inquiry data from the inquiry form, this research intends to make use of pre- and post-inquiry behavioral data by developing an analytics-based framework to extract content profiles from an institutional website and prospect profiles from the institutional server logs. A new sequence of profile-based mass-customized communications was proposed to replace the existing mass communication process. There is a need for models that would help in extracting prospect's behavior, which can be further used in customizing prospective student communications according to the prospect's need. This research intends to address the need through the proposed analytics-based framework.

1.2 Motivation

The application of analytics in higher education industry has been limited to increasing institutional reach, landing page optimization, marketing and increasing traffic to the website. Most of the data collected during and after this process are not being used at all (Bichsel, 2012; Romero, Ventura, Zafra, & Bra, 2009).

With the increased reach, institutions are receiving more inquiries. Due to reliance on models that rely on post-inquiry prospect information from the inquiry forms, institutions are struggling to convert prospects to applicants. This research intends to reduce the inquiry to application gap by focusing on the underlying issues with prospective student communication and engagement.

Studies (Goenner & Pauls, 2006; McCoy, 2011; Noel-Levitz, 2012) clearly stated that most of the inquiries that an educational institution receive are partial inquiries. After extensive research, it was determined that the paths through which the visitors came to make that inquiry and the path through which they left have not been considered as a marketing strategy by the higher education markets. The server logs play a key role in identifying such paths as well as the behavior of different visitors. It captures each and every minute detail of all the navigational characteristics of a prospect (Fischer, Wittern, Schneider, & Tai, 2012; Liu & Keselj, 2007).

Prospects spend a significant amount of time conducting research on the institutions' websites before and after making an inquiry about a program. Using analytics, a prospective student's navigational behavior can be mined from the server logs that record the pages they visited, information accessed, and different actions performed on a page. This pre- and post-inquiry navigational behavior can be used in building prospective student profiles and help us predict the kind of information they expect to hear back from the institutions.

The limitations of post-inquiry data on the existing studies, and the lack of research on prospective student engagement, as well as the application of analytics in a new dimension (admission process) motivated this research to explore the impact of a prospect's pre-inquiry and post-inquiry navigational behavior on the prospect's response rate to the email communications.

1.3 Objective

This research study intends to demonstrate the application of analytics in higher education marketing and address the aforementioned prospect engagement problem. Due to the lack of any existing analytical models in the literature that addresses this prospect engagement problem, it is also the objective of this study to develop an analytics-based framework. From a broader sense, this framework translates the prospect decision-making factors to prospect profiles by identifying the content profiles from the website and prospect navigational profiles from the server logs and mass-customize the communication sequences. The applicability of the framework is then tested in a mid-western public university.

1.4 Outline of Dissertation

This dissertation is organized as follows:

Chapter 2 is a review of existing literature which provides a discussion on the evolution of digital marketing, implementation of digital marketing in higher education, the difference between higher education marketing and e-commerce marketing and further provide a more detailed discussion on prospect decision-making factors.

Chapter 3 provides the research methods and an overall description of the proposed framework, evaluation procedures.

Chapter 4 tests the framework in a higher education setting.

Chapter 5 discusses the results and implications.

Chapter 2 Literature Review

The chapter presents an overview of the existing literature that has a direct bearing on the present investigation.

2.1 Introduction

This research study intersects multiple fields of study including higher education marketing, e-commerce marketing, contextual marketing, web mining, and consumer behavior as shown in Figure 2. Theory base for this research has been extracted from all these areas of research. This section provides the readers with an overview of all the four fields of study and then compares the relevance of these research fields with the ongoing research problem. The following sections provide an overview of domain knowledge i.e., marketing literature in general, how traditional marketing evolved into digital marketing and further elaborate the discussion on analytical knowledge. i.e., data mining and web mining literature.

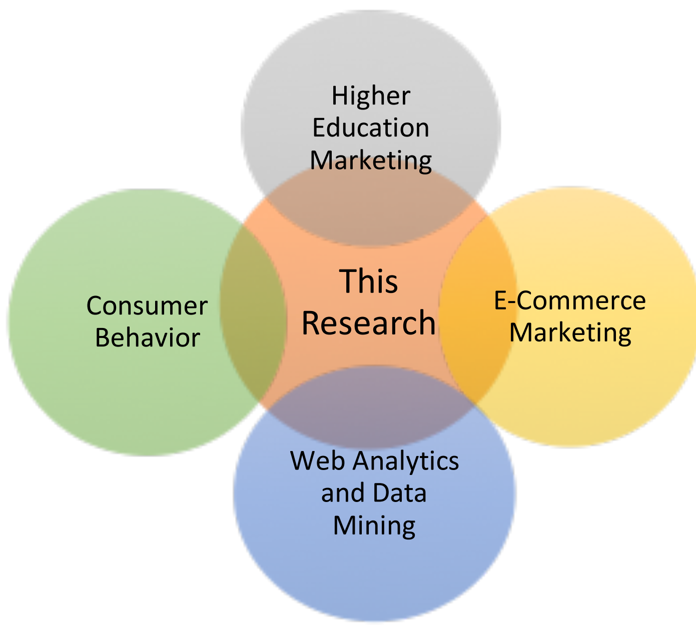

Figure 2: Interdisciplinary Research

2.2 Domain Knowledge

Until late 1980s marketing and advertising in traditional terms made use of newsletters, billboards, flyers, newspapers and print ads. In 1990s the term digital marketing gained the attention of marketing industry (Kiang, Raghu, & Shang, 2000). Later with the evolution of the internet in the 20th-century, people increasingly relied on digital marketing (Corley, Jourdan, & Ingram, 2013; Kiang et al., 2000). People started to make use buzzwords like electronic marketing, online marketing, social media marketing etc. These buzzwords come under the umbrella of digital marketing.

2.2.1 Digital Marketing

Internet became the primary source of information in the 21st century. Search engines began playing an important role in providing reliable information to the users (SEMPO, 2012). Search engine marketing has revolutionized the conventional marketing industry and became a primary source for the marketers to reach numerous people across the globe closing the barriers like time and distance. Interesting results were observed in the 2012 annual survey reports of SEMPO [Search Engine Marketing Professional Organization] as shown in Figure 4 (SEMPO, 2012) clearly depicted the ever-growing market of digital marketing industry in North America. The search engine marketing revenue crossed 26.8 billion U.S. dollars in the year 2013.

Figure 3: The year 2013 search engine marketing revenue report

This clearly depicts the increase in the use of search engine marketing. With the increasing use of search engine marketing, the higher education institutions were able break the geographical barriers in reaching out to the prospects.

The first phase, “Awareness” in the admission funnel discussed in chapter 1 is dependent on the institutional investment in activities such as search engine marketing. Popular search engines like Google, Yahoo, and Bing act as possible interfaces between the institutions and prospects. The institution develops a website with multiple pages containing information about their institutions; educational programs. A site map embedded within the website will help the know-bots and crawlers to easily navigate through the website. The know-bots collect information from the website and index them in the server. The search engines primarily work using crawling, indexing and ranking algorithms (Introna & Nissenbaum, 2000). A simple web search yields a list of web results which are sorted out and placed based on the keyword relevance and page rankings assigned to them by the search engines (Ash, 2008). It is always a challenge for the developers to develop good quality web pages that can withstand the competition, rank first and get placed among the first few pages of the search results (Huang, 2003). In pay-per-click marketing, the search engine identifies the highest bid keyword with good page rank and places it on the first page results. Good sets of relevant keywords are used throughout the text embedded within the web pages. The same set of keywords will be used while placing the ads that will be shown to the users when they make a web search. Users click on a particular search result and are redirected to the advertiser's web page or the landing page (Punera et al. 2010). The landing page should be structured in an attractive way, that it channels the visitors to generate conversions.

Search engines act as a primary source for driving traffic to those websites. Search engines provide thousands of results for every keyword or a phrase and it is observed that users are less likely to browse beyond first few pages, which apparently imply that search engine positioning is highly important for the advertisers to be successful in digital marketing and increase awareness. Richardson et al. (2007) implied that the click-through rate of an ad is directly dependent on the ad position and ads appearing in the latter pages have less visual attention and are less likely to get clicked resulting in the loss of revenue. Getting placed in the search results corresponds to ‘impressions’ and the user action on the search result page i.e. clicking a search result and getting redirected to the web page is a ‘lead’. The user performing any specific action like subscribing or submitting an information request form corresponds to a ‘conversion’. More detailed explanation of different marketing terms like conversions, impressions, clicks, leads, click through rates, bounce rate is explained in Table 1. The marketing terminology that is being used in online advertising is explained in table 1.

Table 1: Basic terminology used in e-marketing

(Grappone et al. 2006; Introna et al. 2000; Jones, 2008; Moran et al. 2009)

Terminology	Definition
Conversion	Specific actions made by the user on the advertiser's website. Examples: submitting an inquiry form, purchasing a product, subscribing to a newsletter.
Impression	Every time an ad appears on a search result or page is counted as an impression.
Click	It occurs when a user clicks the ad leading him/her on the website is a click.
Lead	A visit to a page that enters the conversion cycle and potentially lead to conversion.
Click Through Rate (CTR)	(Number of clicks)/ (number of impressions) i.e. clicks to impression ratio.
Bounce Rate	The percentage of visitors who enters the conversion cycle and exits the web page without making a conversion.

2.2.2 E-Commerce Marketing

E-Commerce marketing or electronic commerce marketing typically deals with marketing and sales of low involvement products and businesses online (Tabibian, Amindavar, & Ave, 2001). Low involvement products involve minimal effort from the consumers before making a purchase and don't have a substantial effect on their lifestyle. Low involvement products may include items that are habitual purchased. Products like a car, house, and higher education program that bring a significant change in a consumer's lifestyle are high involvement products. As shown in the admission funnel discussed in chapter 1, e-commerce marketing has a similar sales funnel structure. E-Commerce marketing in simpler terms relates to providing product related information in the form of e-catalogs and encourages consumers to purchase products (Tabibian et al., 2001). An e-commerce purchase process involves driving traffic to product pages by placing online ads, provide interesting content on the product pages like customer satisfaction reviews about the products and persuade the visitors to purchase the products by increasing awareness (Bhate & Pasha, 2014). Higher education institutions tend to drive traffic to their websites with an intention to provide additional information to the prospect and capture the prospect's information through their inquiry forms. After the inquiry, prospects would receive a series of email communications from the institutions. The following section provides a discussion on higher education marketing and extends the discussion on how higher education marketing is different from typical e-commerce marketing.

2.2.3 Higher Education Marketing

Educational institutions are increasingly relying on digital marketing to build awareness. Products like Google Ad-words, Google Analytics, Yahoo advertising and Bing services are used by educational institutions to promote higher education programs. Unlike e-commerce marketing, higher education marketing becomes more behavioral in nature and the probability of achieving a conversion typically depends on the relevance of the content provided on the page and the visitor's ability to understand the content (Appling & College, 2012). Constructing a solid social community for any specific educational institution is a time-consuming process, which involves quality content to keep the users, engaged. Institutions need to maintain relevant content and an active social presence that engages the members and increases awareness within the community (Gilroy, 2009).

There are different factors that contribute to a prospect's decision in selecting a program in an educational institution. Extensive research has been done in identifying different decision-making factors of prospective students (Moogan, 2011; Moogan et al., 1999; Sheppard, 2013). Moogan (2011) analyzed the decision-making criteria of prospective undergraduates in terms of marketing techniques employed throughout the decision-making period. Sheppard (2013) investigated different factors that influence prospective students in decision-making and the aggregated analysis is provided in table 2. A survey instrument was developed by Sheppard (2013) that addressed six different aspects namely: external influences, education and career goals, information gathering, university financial aid, program characteristics and university characteristics. Noel-Levitz (2012) conducted a survey to extract different factors that influence graduate student college choice. The results varied from online versus regional prospects. Different key prospective student decision-making factors identified from the literature are articulated in Table 2.

Table 2: Key prospective student decision-making factors

Author	Decision-Making Factors
(Aarinen, 2012; María Cubillo, Sánchez, & Cerviño, 2006)	International recognition, suitability, reputation, specialization, quality of the program, courses, future earnings, future job or career opportunities, admission requirements, language requirements, educational facilities, fee, financial aid, City image, institution size.
(Moogan, 2011)	Teaching quality, course content, university reputation, research quality, faculty reputation, accreditation, facilities, student life, career prospects, entry dates, open day, the cost of living, accommodation, friends and family opinion, teacher's opinion, distance from home.
(Morris, 2009)	Electronic catalog, electronic application, inquiry forms, financial aid forms, course registration, email correspondence are some of the key decision-making factors
(Sheppard, 2013)	Program availability, career goals, income, credentials, personal development, flexibility of class scheduling, location, cost of attendance, reputation

The decision-making factors in table 2 have provided this study a foundational reference to extract the content profiles from the website and develop prospect profiles from server logs. The following section provides an overview on the translation of the decision-making factors into prospect profiles, content profiles.

2.2.3.1 Prospect Profiles

Constructing accurate and comprehensive customer profiles play a key role in target marketing and enhanced customer engagement (Adomavicius & Alexander, 2001; Crossley, Kings, & Scott, 2003; Nicoletti, Schiaffino, & Godoy, 2013). In general, profiling can be defined as the recording and analysis of an individual's demographic, psychological and behavioral characteristics (Nicoletti et al., 2013). Building prospective student profiles is a complex task, as prospects do not usually give away explicit information about their interests (Catherine Bounsaythip, 2001; Srivastava, Cooley, Deshpande, & Tan, 2000). Therefore, the prospective student interests must be mined implicitly from the web server logs.

From a prospective student perspective, the general educational purchase process can be described in four phases: general interest in higher education, research for a specific institution or program of interest, decision to apply for one or more schools and finally, making a decision to enroll in a specific program (Goenner & Pauls, 2006). This coincides with the four distinct stages of the higher education admission funnel shown in Figure 1: awareness, inquiries, applications and admissions (Kotler & Armstrong, 2006; Nicolescu, 2009; Noel-Levitz, 2012; Oplatka & Hemsley-Brown, 2004). A prospect in the awareness phase browses through different pages within an institutional website leaving trails of navigational information that can be mined from the server logs. This navigational behavior will be used in the extraction of data and creation of prospective student profiles. Profiling prospective students based on their priorities would help in channeling a prospect to specific communications and increase satisfaction (Bhate & Pasha, 2014). It is proposed that prospective student priorities can be identified from the prospect's existing browsing activities using pre-inquiry and post-inquiry navigational information.

Different decision-making factors identified from literature in section 2.2.3 spoke about the factors consisting of specific contextual information (Aarinen, 2012; María Cubillo et al., 2006;

Moogan, 2011; Morris, 2009; Sheppard, 2013). These decision-making factors are categorized into five different prospect profiles based on their contextual relevance as shown in Table 3.

Table 3: Different prospective student profiles categorized based on targeted information

Prospective Student Profiles	Definition and Targeted Information
Price	Cost of attendance, financial aid availability, cost of living
Program	Availability of the program, online, part-time, distance, location, flexibility of class schedules, distance from home, city image
Future employment	Career advancements and goals, course content, future jobs, future earnings, on-campus employment, income credentials
Institutional Image	Institutional Reputation, teaching quality, faculty expertise and reputation, research quality, quality of the program, institutional size
Environment	Technology use, educational facilities, student life

Institutions provide specific information addressing these decision-making factors on their websites. Based on these prospect profiles, the web pages within the institutional website are categorized into different content profiles. Pages that address a specific context are tagged with prospect profile names based on the context as followed: Price, Program, Future Employment, Institutional Image and Environment. Contextual grouping identifies contextual web content then categorizes web pages based on the prospect profiles derived from research as indicated in Table 3. Using the contextual groupings, sequences of mass-customized communications are developed based on prospective students' pre- and post-inquiry navigational behaviors.

2.2.3.2 Importance of Prospect Profiles

This section provides an overview of crucial decision-making factors and their importance in this research. Extensive study has been conducted on identifying different influential factors that impact a prospect's choice of an institution or program. Ivy & Naude (2004) introduced a 5P model where the 5P's stands for product, price, place, promotion and people. In a detailed sense, the product represents the program a prospect is going to invest; price represents the cost of the program, place deals with the environment or location of the institution, promotion targets the future employment and people deals with the student life and institutional image.

Filip (2012) proposed a 7P model and included processes and physical facilities to the existing 5Ps. Processes refer to the way the enrollment system, teaching and learning habits, social and sports activities are established within the institution. Physical facilities address the institutional equipment, technical infrastructure etc. Although it is up to an institution to target specific factors in reaching out to prospects, a research study conducted by Noel-Levitz (2012) clearly articulated that cost, financial aid, academic reputation, institutional size, future employment and campus location are crucial prospect decision-making factors.

Higher education institutions structure their program-related information targeting these specific decision-making factors. To my knowledge based on the conducted literature review, the existing research didn't provide any models that have addressed the 7p's and translated them into a prospect engagement strategy.

2.2.3.3 Need for customized communication sequences

During the inquiry phase, higher education institutions use various types of promotion including emails, printed materials, and campus visits as different means to reach out to prospects. Among all these modes of communication, emails play an important role. Gomes & Murphy (2003) elucidated that over 80% of the prospective students consider the email communications received from educational institutions as an influence in their choice of selecting an institution. In general, a prospect may receive the following information in the form of emails: the cost of the program, availability of the program, alumni testimonials to motivate the prospects to apply, and a message from the dean or program coordinator.

Morris (2009) addressed the relation between timeliness of contacts and likelihood of getting a prospect enrolled for a program. He articulated that the number of days from the date of inquiry and date of application, the number of days between the date of application and date of admission, and the numbers of student-initiated contacts with the institution are strong predictor variables in enrolling a prospect. Naidoo & Connect (2011) argued that the content a prospective student receives during this timeframe has a direct impact on the prospect's initiation to stay in contact with the institutions. The existing research made use prospect information like student demographics and household income that are extracted from the applications to predict the enrollments (Hemsley-Brown & Oplatka, 2006; Moogan, 2011).

Institutions are increasingly relying on business intelligence tools and customer relationship management (CRM) concepts to achieve a competitive advantage (Labus & Stone, 2010). In CRM the ability to provide content and services that are tailored to the individual's preferences is an important marketing tool (Phan & Vogel, 2010). Adomavicius & Alexander (2001) stated that tailored and personalized communications impact the organization's return on investment in a positive way. Communicating with a prospective student and keeping the prospect engaged with the institution until the end of the admissions funnel is an arduous process. Prospects' tend to drop out from the admission funnel if they don't receive relevant information from the institutions (Noel-Levitz, 2012).

Studies indicated that a prospective student usually applies to more than three educational institutions and it is highly important to craft the communications so that the institutions might not lose the prospect within the admission funnel (Sheppard, 2013). Gomes & Murphy (2003) & Aarinen (2012) articulated that prospects experienced dissatisfaction during the pre-application phase due to difficulty in finding information on the website or receiving irrelevant information through the communications they received from the institutions.

Morris (2009) analyzed the impact of prospective student interaction with the college and engagement on the enrollment by linking the timeliness of the communications and in-turn the interactions to the likelihood of student enrollment. The study findings confirmed that timeliness and time intervals of interactions are successful in predicting the purchase likelihood.

The following section will discuss the student engagement strategy, the importance of subject lines and their impact on email open rates.

2.2.3.4 Student Engagement Strategy

Email communications battle other media in an attempt to capture the prospect's attention. As such, it is critical that the performance of email communication process is refined. The email communication performance can be measured by the following email marketing metrics: Click-through rates, open rates, and deliverability rates (Zarrella, 2014).

Open rates play a crucial role in measuring the performance of email communications as it begins the engagement process. POP (2015); Watson, Pay, Fidura, & Quist (2013); Zarrella (2014) articulated that the average email open rate is around 18% and the average click through rate is around 9% for the higher education markets. POP (2015) stated that open rate is a best metric to track the progress or to identify problems with engagement. Watson, Pay, Fidura, & Quist (2013) articulated that the email marketing has three stages of engagement namely: Interest (subscribe), Attention (open emails) and Action (click on the link). Translating the same into higher education marketing, interest relates to a prospect making an inquiry about a program, attention relates to a prospect opening and reading an email communication from the institution and action relates to click on the link in the email or make a decision to apply. A prospect may apply at any stage of the engagement cycle and the following figure clearly illustrates the three different stages interest, attention, and action. As discussed in the admission funnel the awareness phase creates interest within the prospect, the inquiry phase contributes to the attention and action, finally, the application phase can be related to conversion and sales.

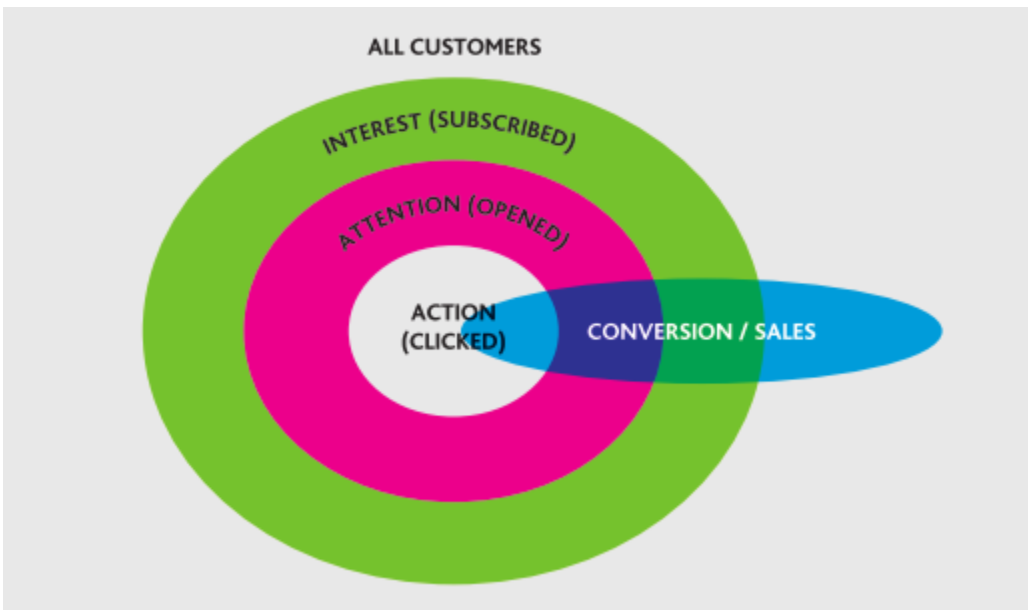

Figure 4: Stages of engagement in Higher Education Marketing

2.2.3.5 Impact of subject lines on open rates

This research uses open rates to analyze the performance of a communication sequence base on navigational behavior and contextual groupings based on prospect profiles. As such, an understanding of subject lines and their impact on open rates is critical.

There is no one-size-fits-for-all solution to determine the length of the subject line, but research has shown that prospects tend to open emails with precise, shorter, relevant and personalized subject lines (POP, 2015; Watson et al., 2013; Zarrella, 2014). As the characters in the subject line increase, the studies indicated a significant decrease in open rate as well as the click-through-rates of email communications (POP, 2015; Zarrella, 2014).

The number of emails people receive has increased exponentially and they started prioritizing the emails for attention over others. Most people tend to spend a fraction of a second evaluating email subject fields before moving on to the next. Extensive research has been done on the email surface features that draw the recipients attention to it (Wainer, Dabbish, & Kraut, 2011). The subject lines play a crucial role in the open rate of emails, especially when they are viewed on mobile devices. Considering the existing email marketing literature that stated that around 48% of the emails are opened on mobile devices and 69% of the mobile users mostly delete the emails that are not optimized for mobile devices (Zarrella, 2014).

Email open rate indicates the increased institutional reach to the prospects. This would eventually lead to the increase in prospect engagement. The subject lines play an important role in capturing the prospect's curiosity and later persuade him/her to perform an action on the email received. It is estimated that subject lines with 50 characters or less result in 12% higher open rates than that of the emails with longer subject lines. This research study intends to test the prospect engagement by capturing the open rates for the mass communication sequences to that of the mass-customized communication sequences.

The following sections will provide more details on existing communication plans and the proposed mass-communication plans.

2.2.3.6 Existing Communication Plans

In a report, Noel-Levitz (2012) stated that in most of the private and public universities, pre-structured mass communication plans are being used for engaging the prospective students. The existing communications sequences are structured based on the prospect's post-inquiry information obtained from the inquiry forms. Such communication system will not deliver relevant information to the prospects when the institution receives partially filled inquiries.

The following section provides an overview of the analytical knowledge that is used in this research, and certain concepts of data mining and web mining are used to extract content profiles from the website and prospect navigational profiles from the server logs.

2.3 Analytics Knowledge

With the evolution of the web, institutions started collecting large amounts of information from consumers. The web servers collect and store visitor information whenever a user visits a website. E-Commerce marketers were able to make use of such data to provide product recommendations to consumers and engage with them throughout the purchase process (Kelly, 2003). They even became successful in using the information to mine the consumer behavior and apply that knowledge to remarket themselves to potential prospects'. The following sections will provide an overview of data mining and web mining concepts that are used in deriving such knowledge and how they will be applied to this research.

2.3.1 Data Mining

Data mining is discovering knowledge from a database. It is a process of extracting implicit, non-trivial, previously unknown and useful information from collected data (Chen, Han, & Yu, 1996). Techniques like neural networks, decision trees, and genetic algorithms are different categories that come under the umbrella of data mining. Neural networks evolved on a quest to predict visitor behavior. Decision tree algorithms are used for visitor comprehension, genetic algorithms for visitor categorization and k-nearest neighbors for segmenting visitors (Kelly, 2003). Customer segmentation is best achieved by application of different clustering techniques'. Expectation-Maximization(EM) is a clustering technique that is used in segmentation problems. This research extracts prospect sessions from server logs and cluster prospective student behavior using EM technique. The following section describes the concepts of web mining that will help this research in extracting the prospective student behavioral clusters.

2.3.2 Web mining

The application of data mining techniques to World Wide Web is web mining (Bamshad Mobasher, Namit Jain, Eui-Hong Han, 1997). Although the concept of web mining has been in place for over a decade. Web mining is still an emerging topic and lot of industries are still trying to adapt and implement web mining (Vellingiri & Pandian, 2015). Srivastava et al., (2000) articulated the taxonomy of web mining into two distinct categories, the web content mining, and the web usage mining. Later a third category named web structure mining is also added to the taxonomy (Kosala & Blockeel, 2000). Web content mining is extracting contextually relevant web pages by mining the underlying content, whereas the web usage mining addresses the problem of analyzing user navigational and behavioral patterns from the server logs (Catherine Bounsaythip, 2001). Web structure mining, on the other hand,is a knowledge extraction process from the structure of hyperlinks and underlying HTML documents (Vellingiri & Pandian, 2015). This research intends to make use of web content mining to extract the contextual profiles of the web pages within the website and web usage mining to extract the prospect profiles from server logs and compare them to determine the mass-customized communication sequences.

This study intends to address the research gap by developing a framework that translates the prospect decision-making factors into prospect profiles. Different prospects will have different needs and one communication sequence will not address all the prospects needs (María Cubillo et al., 2006).

In order to provide relevant information to the prospect's, his/her interests have to be extracted from the web server logs. Instead of a mass communication sequence to all the prospective students, this research intends to mass-customize the communication sequences based on the prospect profiles.

This research attempts to apply content mining techniques to extract the content profiles from the website and make use of web mining and data mining techniques on the server logs to create prospective student profiles, map prospective student profiles to the content profiles and identify mass-customized communication sequences. The email open rates for the mass-customized communication sequences and the mass communication sequences will be compared in the further sections.

Chapter 3 Methodology

3.1 Introduction

This chapter provides an overview of different objectives of conducting this research and different research methods that are applied throughout the study.

3.2 Research Objectives

Given the rich visitor navigational information available from the web server logs, it is argued that the pre-inquiry and post-inquiry navigational data will provide much richer insights in identifying the prospective student profiles. The server logs comprise of the navigational activity on the website. Prospect profiles will be created from the existing server logs. These prospect profiles will be used to mass-customize the communication sequences. To the author's knowledge, no study has been conducted that has utilized the pre-inquiry and post- inquiry navigational behavior of a prospect that would impact the prospect's institutional choice.

This dissertation study has the following objectives:

- Develop a framework for identifying prospective student profiles by utilizing user navigational patterns
- Suggest mass-customized communication sequences based on the content profiles identified from the website and prospect navigational profiles identified from server logs,
- Validate the subject lines with its contextual relevance to the identified prospect profiles.
- Validate the framework and communication sequence by practical implementation of the new sequence of communications and track the respective open rates to the communications.

3.3 Research Methods

This research intends to apply design science research methods for developing the framework. Design science research principles proposed by Hevner, March, Park, & Ram (2004) and Peffers, Tuunanen, Rothenberger, & Chatterjee (2007) act as a guideline for the construction of the framework as shown in the design science research methodology process model and is explained below.

Design science or design research can be defined as an explanation about the improvement on the behavior of aspects of information systems and analyze the performance of the designed artifacts (Hevner et al., 2004). Design science research recommends the building and evaluation of software artifacts. It is concerned in devising artifacts to attain goals that may serve human purposes (Peffers et al., 2007).

Design research methodology follows a systematic approach to improving artifacts and behavior of aspects in information systems (March & Smith, 1995). This research methodology mainly focuses on increasing the need for understanding the phenomena of design methods and tools that helps to improve the observed situations in design (Hevner et al., 2004). Design research methodology is not a sequential process, several parallel stages will run and many iterations may take place before advancing to the state of the art (Peffers et al., 2007). Hevner et al., (2004) proposed seven design science research guidelines that will channel a researcher to develop an information technology artifact. This research study adheres these seven guidelines:

- Design as an Artifact
- Problem Relevance
- Design Evaluation
- Research Contributions
- Research Rigor
- Design as a Search Process
- Communication of Research

Design as an artifact can be defined in such a way that a purposeful IT artifact is created to address an important research problem (March & Smith, 1995). The artifact should be effectively described, enabling its implementation on an appropriate domain (Hevner et al., 2004). The objective of the Design is to develop applications or technology-based solutions that are relevant to business problems (Hevner et al., 2004).

Evaluation plays a key role in the research process and well-executed evaluation methods will demonstrate quality efficient design artifacts (Peffers et al., 2007). The IT artifacts are typically evaluated in terms of functionality, completeness, reliability, usability etc. (Hevner, 2004). The research rigor addresses the way in which research is conducted. Design science research requires rigorous methods in both construction and evaluation phases of the artifact design (Hevner, 2004).

The researcher's search and formulate the optimal design that is often adopted by real-time information system problems (Peffers et al., 2007). According to Peffers et al., (2007) the design science research process follows six specific steps: Identify problem and motivation, the objective of a solution, design, and development, demonstration, evaluation, and communication. The Sample design science research is explained in the six stages: awareness of the problem, objectives, design & development, evaluation and communication as described below in Figure 5:

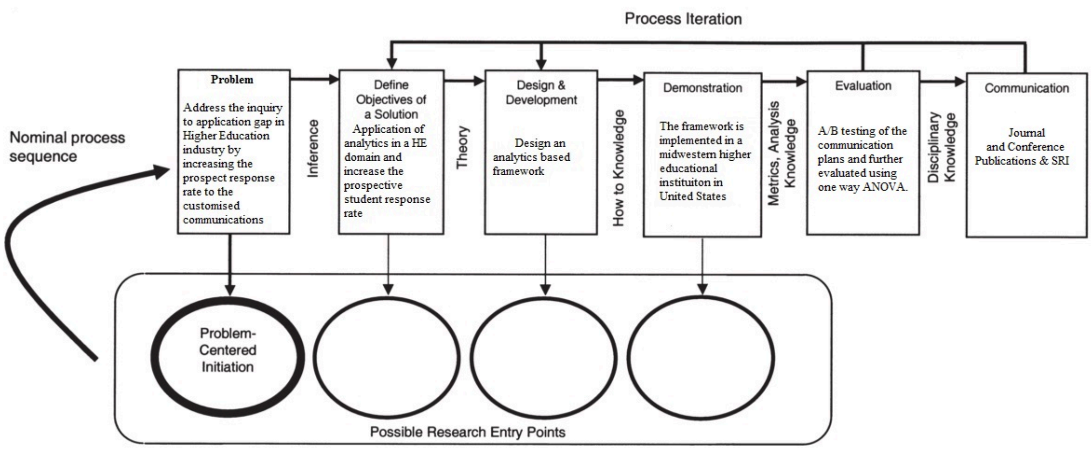

Figure 5: Design science research methodology

3.3.1 Problem

The process of identification and motivation of a problem requires the researcher to demonstrate the need for that model, which will be motivated by problem-centered initiation (Peffers et al., 2007). This research problem intends to address the inquiry to application gap from a broader perspective. Higher education institutions tend to lose over 75% of the prospects in between the inquiry and application phases of an admission funnel and the application of data mining techniques in higher education institutions are only limited to reporting purposes (Ali, 2013). This research intends to expand the use of data mining techniques by extracting the content profiles from the website, link it with the prospect navigational behavior observed from the server logs and generate the mass-customized communication sequences. These mass-customized communication sequences were tested against the existing communications sequence and the significant difference in the email open rates were observed.

3.3.2 Objectives of Solution

Objectives of the solution include the need to identify an artifact, which is motivated by objective centered solution (Hevner et al., 2004; Peffers et al., 2007). The objective of the solution for this research is derived from the literature review section as follows:

- Address the inquiry to application gap by making use of profile-based, mass-customized communication sequences.

3.3.3 Design and Development

Development involves the process of developing the framework, which will be motivated by a design and development centered initiation (Hevner et al., 2004; Peffers et al., 2007). The proposed framework consists of four different phases: web & text mining, pre-processing, clustering and profiling, and customization that are described as followed:

1) In the web and text mining phase, all the web pages within the institutional website are processed and the web pages are categorized into content profiles based on the contextual information.
2) In the pre-processing phase, all the server logs are processed to extract individual sessions from the navigational information in the form of pre-inquiry, inquiry and post-inquiry phases.
3) Prospective student profiling will be achieved through clustering and segmentation. In the clustering and profiling phase, prospective student segments are clustered into specific groups by investigating the EM-clustering technique to identify the maximum likelihood of a group of prospective students that come under a specific profile.
4) In the customization phase, the popular sequences based on the frequency of visits obtained to each page in each cluster will be used to determine the mass-customized communication sequence.

3.3.4 Demonstration

The demonstration is the process in which the developed artifact is implemented to solve the problem. This step requires an appropriate context or a client where the artifact can be implemented (Hevner et al., 2004; Peffers et al., 2007). The proposed artifact is demonstrated in a public university in mid-west of United States on the prospective graduate student population. The effectiveness of the email subject lines will be validated through a pilot study and the mass-customized email communications will be validated through controlled field testing. More detailed explanation of the demonstration is provided in the following Chapter 4.

3.3.5 Evaluation

Evaluation step involves the process of analyzing the ability and productiveness of the artifact, based on which the researcher will think if he/she should proceed with their artifact or should revisit the design step (Hevner et al., 2004; Peffers et al., 2007). A control group and a treatment group are considered for the evaluation of this study. Predefined mass communications are sent to the control group and profile based mass-customized sequences of communications are sent out to the treatment group. The chi-square test is conducted and the control and treatment groups will be evaluated for statistically significant results. More detailed explanation of this process is provided in the following section 3.5.

3.3.6 Communication

The knowledge acquired through the evaluation process is to be shared with fellow researchers through scholarly publications and other journal resources (Hevner et al., 2004; Peffers et al., 2007). More detailed explanation is provided in chapter 5.

3.4 Framework

In this data-driven environment, higher education institutions tend to collect huge amounts of prospect information from different sources (Ali, 2013). Much of this data is self-reported and includes general information about the program of interest, anticipated entry date and how the prospect came to know about the program. Unlike self-reported information, server logs collect specific information regarding pages navigated throughout the website, time of entry, time of exit and time of conversion. This research specifically deals with the information extracted from the server logs and the use of the data in the framework. The following sections provide the underlying theoretical foundation for the framework.

3.4.1 Theoretical Foundation

Desjardins (2002) implemented an analytical strategy to assist higher education institutional marketing efforts. Desjardins (2002) applied a conceptual model based on the human capital theory that considered variables like current student demographics, admitted years, enrolled programs, application forms etc. Desjardins (2002) attempted to fit a statistical model by considering the historical data of admitted students and tried to accurately predict enrollment. Following this work, Goenner & Pauls (2006) proposed a model to predict the enrollment decisions of prospective students based on their inquiries. Goenner & Pauls (2006) combined the prospective student demographics with US census data and proposed that the prospects from a specific geographic region behave in a specific pattern.

Goenner & Pauls (2006) & Desjardins (2002) predicted a prospect's enrollment decision and then suggested specific marketing communications channels for prospective students, current students, and alumni. Moogan (2011) specified that customer specific information in the communications chain might improve the retention rates as well as the brand image of the institution. He also articulated that the prospects are to be considered as valuable customers and complete effective communications are to be exchanged between the prospects. Such relationship needs to be established from the inquiry phase till his/her graduation from the program. According to Moogan (2011), most of the existing research was conducted before or in the early stages of the evolution of online marketing and do not reflect the current marketing technologies to match the information needs of the students. This research attempts to generate a dialog on the higher education marketing efforts and then contribute to the institutional advancement by making use of analytics.

A theoretical model shown in Figure 6, is derived from the works of Desjardins (2002); Goenner & Pauls (2006); Hossler & Gallagher (1987); Michael (1990). Hossler & Gallagher (1987) articulated that a student's choice of an institution depends on the predisposition of pursuing higher education, selecting schools of interest and apply to the institutions based on choice. Desjardins (2002) articulated that higher education institutions tend to make use of economic models and business intelligence models in promoting marketing methods and reporting. Goenner & Pauls (2006) developed a model that made use of the prospect demographic and financial information extracted from the applications and predicted the enrollment numbers.

Michael (1990) emphasized on different factors that influence the prospects behavior from choosing one university over the other.

The theoretical model illustrates the interdependency of the student choice in researching for potential schools in the market as well as the institutional efforts in capturing the prospective student's attention through their communications. This prospective student information can be further mined and used to personalize and customize the communications that have a direct impact on persuading prospective students. According to Oinas-kukkonen & Harjumaa (2009), information tailored to the potential needs of a prospect or to the interests and personality of a prospect will be more persuasive. From the theoretical model, it is clear that student choice, institutional efforts and the communications that a prospect receives will persuade a prospective student's decision-making behavior.

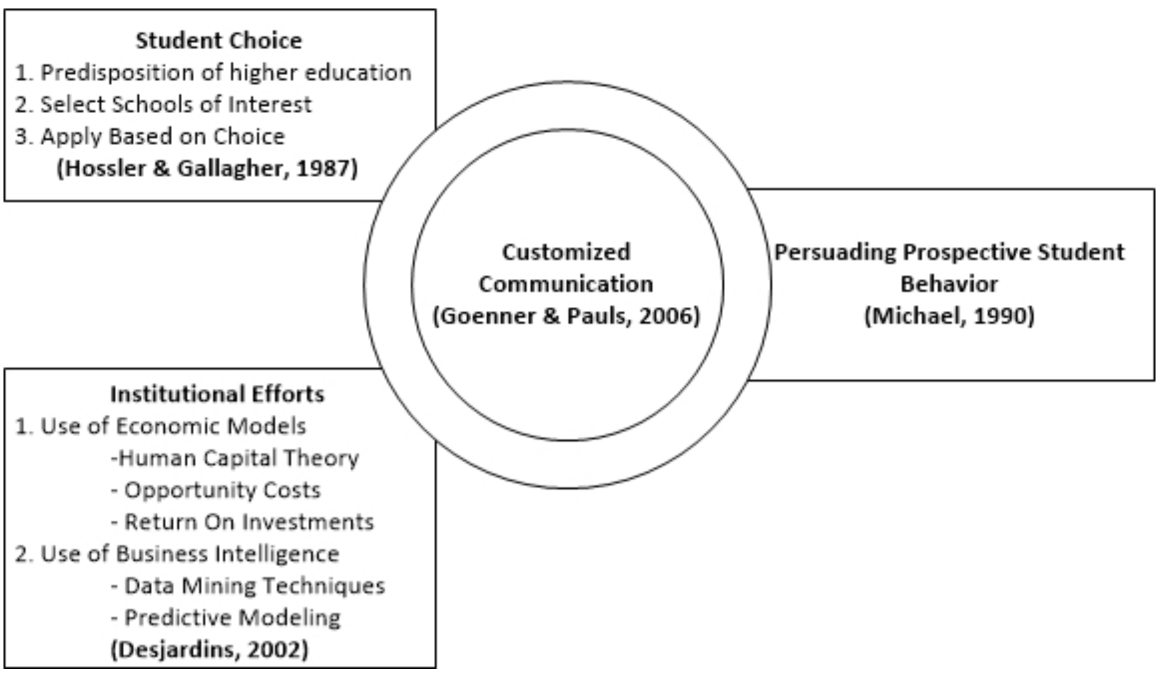

Figure 6: Theoretical model derived from literature

Liu & Keselj (2007) articulated that key behavioral aspects of a visitor can be identified from actual server logs. They identified that frequency and duration are two key factors that typically represent a user's interest in the content on a page. If a page has a higher bounce rate, it indicates that the page might be carrying irrelevant content, which needs some attention. A bounce rate is captured when a user leaves the advertiser's website without making any specific action. Liu & Keselj (2007) also indicated that the sequence of pages that were visited by a visitor could also be considered while analyzing users' preferences and interests. The number of times a page has been visited, time spent on the page and the number of visitors who visited the same page before making an inquiry are some of the factors that determine the visitor behavior. The following section will expand more on the design of the framework and the underlying phases.

3.4.2 Analytics-Based Framework

The framework consists of four phases as shown in the following figure 3:

1) Web & Text mining phase:

   In the web & text mining phase, content from each page within the website will be extracted and text mining techniques will be used to extract the content. All the pages that speak about a specific context will be extracted and grouped into content profiles in this process. All the pages under each profile will be related to the prospect profiles identified in table 2.

2) Pre-processing phase:

   In the pre-processing phase, individual prospect student sessions containing visitor navigational information will be extracted from the server logs. Each session consists of individual prospect level pre-inquiry, inquiry, and post-inquiry information.

3) Clustering and Profiling phase:

   In the clustering & profiling phase, unsupervised clustering techniques will be applied to the prospect sessions extracted in the pre-processing phase. Each and every page will be assigned to a cluster. Each cluster will be assigned to the prospect profiles identified in table 2. The prospect profiles contain a set of most frequent web pages accessed by each group.

4) Customization phase:

   In the communication and relationship management phase, prospect profiles identified in the clustering phase will be mapped to the content profiles identified in the web and text mining phase. The existing mass communications will be re-sequenced based on the popular or sequence of frequent pages accessed under each cluster.

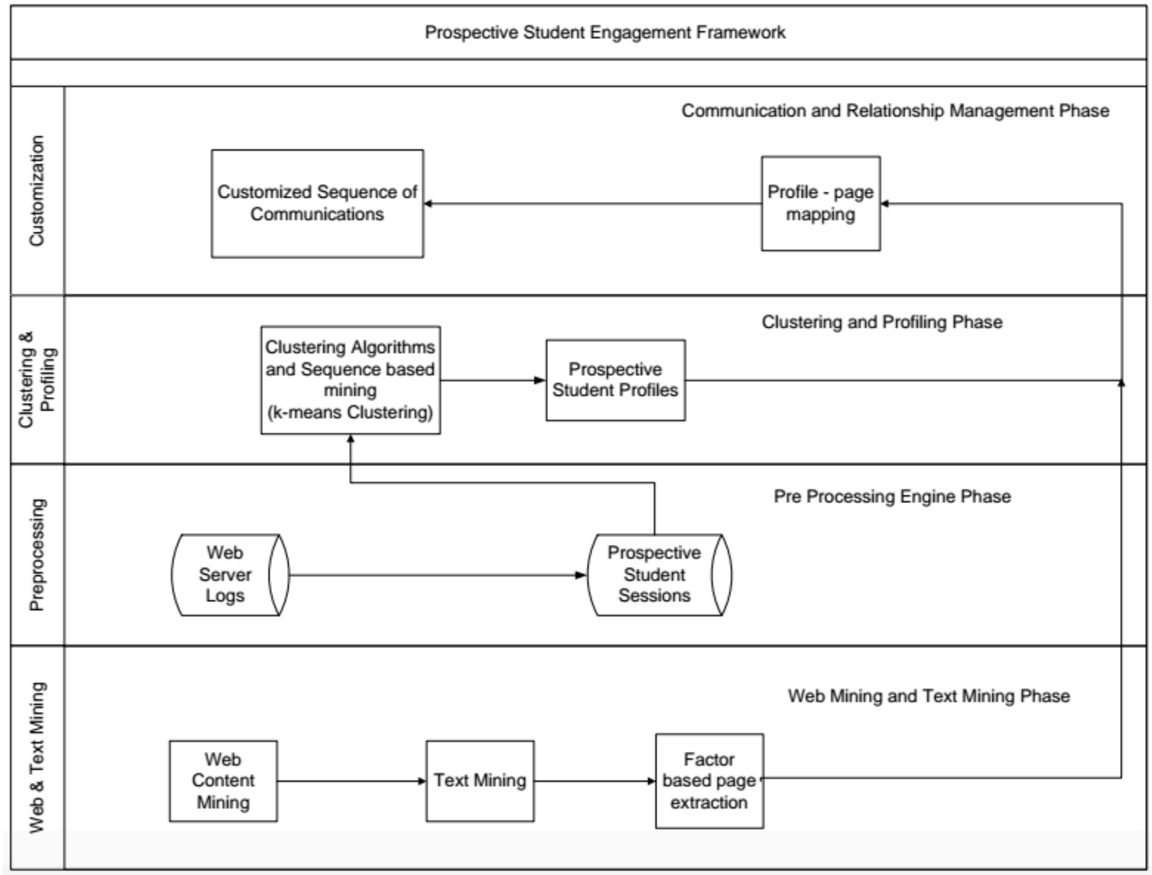

Figure 7: Analytics-based framework

3.4.2.1 Web & Text Mining Phase

The main objective of this phase is to extract the content profiles from the institutional website. In this phase, the underlying web page content is extracted with the help of a crawler. By the application of text mining concepts such as term identification, stopword detection, case transformation, and n-gram detection, different attributes are identified that have a contextual relation to the decision-making factors identified in Table 1.

This phase will allow the researcher to link the information provided on the institutional website to the decision-making factors extracted from the literature and to the prospective student profiles. The web mining and text mining phase will essentially crawl the institutional website, mine the content on all pages, tokenize the words in the content and group similar web pages that talk about a similar context within the website (Husin, 2013; Markov & Larose, 2007). These groups will be mapped to the decision-making factors discussed in table 2.

The institutional site map will act as an input for the web mining and text mining phase. The content from all the pages within the institutional website will be extracted and the text-mining concepts will be applied to the content to categorize the pages that speak about a specific context. A group of pages that have information about a specific context will be identified and contextual profiles will be extracted from this phase. These contextual profiles will help the researcher in determining the information used in communication plans relate to the decision-making factors.

3.4.2.2 Pre-processing Phase

The main objective of this phase is to extract individual sessions that contain the pre-inquiry, inquiry, and post-inquiry information. The pre-processing phase deals with the data collection, cleaning the outliers and information extraction tasks. The data collection involves tracking visitor navigation to an institutional website via server logs. The researcher currently has access to IIS-6 server logs from a mid-western higher education institution in the United States. These server logs are a set of records containing client and server side communication. Whenever client computer intends to access a web page, the client computer will send a request to access the web page from the server computer. After the client computer's request, the server computer will record the client computer configuration about the type of browser, version number etc. from the client computer. Each record in the server log will contain multiple instances of client and server communications that act as outliers.

The raw data set is the input of this phase. The raw data set consists of all the unnecessary client-server communication that are deleted in this pre-processing phase. Individual sessions that consist of pre-inquiry, inquiry and post-inquiry information are extracted from the server logs. Based on the client accessed, the web page and time the pre-inquiry, inquiry and post-inquiry phases are determined.

A server log comprises of different kinds of variables like date, time, server site name, server IP, client side method, client accessed URL, client-side query, server port, client username, client IP, client-side user agent, server status, server sub status and server windows-32 status as shown in figure 6. From all these variables very few variables like date, time, client IP, and client accessed URL are key to this research. This section will provide an overview on how these key variables are related to the extracted profiles.

date	time	s-sitename	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port	cs-username	c-ip	cs(User-Agent)	sc-status	sc-substatus	sc-win32-status
9/1/13	0:00:06	W3SVC1814367467	#	GET	/emergency/management/Public/RSS.aspx	-	80	-	#	-	200	0	0
9/1/13	0:00:06	W3SVC1814367467	#	GET	/includes/sliders/homepage-slider.aspx	-	80	-	#	-	200	0	0
9/1/13	0:00:06	W3SVC1814367467	#	GET	/rss/home.xml	-	80	-	#	-	200	0	0

Figure 8: Sample server logs

Date, time, client IP and client accessed URL are interlinked variables. Each row in the sample shown in figure 6 represents an instance or a click from a visitor. Date and time provide data regarding the day and time the instance has occurred. Client IP is a distinct variable that represents the prospect. Every unique IP address is considered as an individual prospect. Client accessed URL represents the specific page or information accessed by the prospect from an institutional web page. The institutional web pages provide information related to the prospect decision-making factors discussed in table 2. Based on the date, time of the inquiry and client IP address critical distinction of pre-inquiry, inquiry and post-inquiry phases of an individual session are extracted. These sessions are clustered further to extract the sequences.

3.4.2.3 Clustering and profiling Phase

The main objective of this phase is to extract the prospect navigational profiles from the pre-processed data. The clustering and profiling phase will test unsupervised clustering techniques to the pre-processed data to find the best fit model. In this case, k-means clustering technique is applied and prospective student profiles are generated from the clusters. K-means clustering is an unsupervised learning technique that classifies a given data set into a certain number of pre-determined clusters. The clusters generated through this model are not evenly distributed so the Expectation Maximization model is applied. From the literature, the prospective student profiles are categorized into five different categories price, program, future employment, institutional image and environment as shown in Table 3. These five categories are considered in determining the number of clusters to be extracted from the data set and as the number of clusters are predetermined this research tend to apply K-means clustering technique. In this profiling phase, we relate the clusters to the existing student characteristics that are already identified and grounded in higher education marketing literature (Kathuria, Jansen, Hafernik, & Spink, 2010; Song & Shepperd, 2006). The clusters consist of a sequence of pages consisting of prospect clicks for a specified timeframe. These clusters will support the prospective student profiles' best fit.

The individual sessions identified in the pre-processing phase will act as inputs for this phase. The sessions consist of pre-inquiry, inquiry and post-inquiry navigated web pages. Through the application of EM clustering technique, different clusters with frequently accessed pages were extracted. The frequency of visits to pages in each cluster will help in prioritizing the communication sequences in the next step.

3.4.2.4 Communication and Relationship Management Phase

The main objective of this phase is to extract the mass-customized communication sequences based on the content profiles and prospect profiles. The communication and relationship management phase will customize the sequence of communications based on the individual prospective student needs.

By visualizing the profiles identified under each cluster, pages with a higher frequency within the cluster are identified and prioritized. The web pages in each cluster are prioritized based on the frequency of visits. These pages are then compared with the contextual profiles and the new sequence was identified from each cluster. Although popular pages can be extracted from other sources, such information is not limited to only prospective student information. Through the identification of sessions containing pre-inquiry, inquiry, and post-inquiry information confirms that this information is specifically limited to prospective students. Each profile will have a set of communications that are prioritized based on the frequency of visits the web pages received within the cluster. The existing mass communications consist of a single sequence of messages that are sent out to prospects based on the anticipated entry date and irrespective of the prospective student needs. Since a single sequence of mass communications will not address all the prospects needs, this research study proposes to make use of profile-based sequencing and mass-customized communication sequences instead of mass communications. These mass-customized communications will help with the following:

[1] Customizing information to the prospects,

[2] Increasing the engagement and institutional awareness with the prospects and,

[3] Establish, enhance and retain a relationship with the prospects.

The prospective student profiles extracted from the clustering and profiling phase and the content profiles identified in the web and text mining phase will be used as inputs for the communication and relationship management phase. The content profiles obtained from the web and text mining phase will be compared to the navigational behavior obtained from the clustering and profiling phase extracting the new sequence of mass-customized communications for each prospective student profile.

3.4.3 Privacy Concerns

The Midwestern public higher education institution selected for this study complies with Family Educational Rights and Privacy act (FERPA). The privacy policy supports the privacy of information submitted by all students, faculty, staff and visitors who visit the official website (University, 2015). In the privacy policy, it is clearly stated that the institution will collect navigational information such as IP addresses, browser information, search terms, browsing patterns for analytical and statistical purposes and this does not include any personal information linked to any particular individual. The data is cumulative and individuals are unidentifiable (University, 2015). The information is used to address any technical problems, provide more relevant information to users and other administrative functions. This research inclines towards providing relevant and customized information to the prospective students.

3.5 Evaluation

This objective of this section is to describe the formative and summative evaluation for this research study. The formative evaluation is achieved by considering the experimental setting and assessing needs, and analyzing the existing data sources. The summative evaluation is achieved by formulating hypotheses and statistically validating them. The following section will provide an overview of the summative evaluation.

3.5.1 Hypotheses for the evaluation of communication sequences

Through this experiment, this research intends to test and verify the response from prospects in terms of open rates to the email communications they receive from the institutions with a change in communication sequences.

Null Hypothesis - H_0: The likelihood of a response from a prospect to the email communications remains same regardless of the change in the sequence of communications he/she receives from the institutions.

Let "x" be the response rate to the email communications and "mu" (μ) be the population of the control group.

$$H_0: \mu = X$$

Alternative Hypotheses:

- H_{a1}: The likelihood of a response from a prospect to the mass-customized email communications increases with the change in the sequence of communications he/she receives.

$$H_{a1}: \mu > X$$

- H_{a2}**:** The likelihood of a response from a prospect to the mass-customized email communications decreases with the change in the sequence of communications he/she receives.

$$H_{a2}: \mu < X$$

3.5.2 Hypotheses for the evaluation of the influence of context

Through this experiment, this research intends to test and verify the response from prospects in terms of open rates to the email communications they receive from the institutions with a change in the contextual preference.

Null Hypothesis - H_0: The likelihood of a response from a prospect to the email communications remains same regardless of the change in the sequence of communications he/she receives that matches the contextual preference of the prospect.

Let "x" be the response rate to the email communications and "mu" (μ) be the population of the control group.

$$H_0: \mu = X$$

Alternative Hypotheses:

- H_{a1}: The likelihood of a response from a prospect to the mass-customized email communications increases with the change in the sequence of communications he/she receives that matches the contextual preference of the prospect.

$$H_{a1}: \mu > X$$

- H_{a2}**:** The likelihood of a response from a prospect to the mass-customized email communications decreases with the change in the sequence of communications he/she receives that matches the contextual preference of the prospect.

$$H_{a2}: \mu < X$$

These hypotheses were tested in a public university setting. The following section provides a detailed discussion on the evaluation.

Chapter 4 Research Findings

This research made use of server log data from a mid-western public university in the United States.

4.1 Data Collection

This study made use of server logs containing visitor information from a public university in Midwest, United States. Initial tests were conducted on the data set for the year 2013. These raw server logs were a collection of all the visitor information visiting the university website. These log files contain unnecessary client-server communications. As shown in the below Figure 9, the IIS 6 server logs from the university contains Client IP address, Username, Date, Time, Service and instance, Server name, Server IP address, Time taken, Client bytes sent, Server bytes sent, Service status code (A value of 200 indicates that the request was fulfilled successfully.), Windows status code (A value of 0 indicates that the request was fulfilled successfully.), Request type, Target of operation, Parameters (the parameters that are passed to a script). The raw server logs comprised of over one million records. These records include client-server communication, on-site activity as well as prospect inquiries. The raw server logs were cleaned in the pre-processing phase and individual sessions are extracted as discussed in the below section 4.3.

date	time	s-sitename	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port	cs-username	c-ip	cs(User-Agent)	sc-status	sc-substatus	sc-win32-status
9/1/13	0:00:06	W3SVC1814367467	#	GET	/emergency/management/Public/RSS.aspx	-	80	-	#	-	200	0	0
9/1/13	0:00:06	W3SVC1814367467	#	GET	/includes/sliders/homepage-slider.aspx	-	80	-	#	-	200	0	0
9/1/13	0:00:06	W3SVC1814367467	#	GET	/rss/home.xml	-	80	-	#	-	200	0	0

Figure 9: Raw server logs

4.2 Phase-1

In this phase, all the existing URL's in the institutional website are extracted by making use of a crawler, then text mining rules are applied on the web pages. Through the web content mining, the underlying web page content is extracted. Term identification, stopword detection, case transformation, and n-gram detection techniques were applied to contextually profile the web page content. Different attributes are identified that have a contextual relation to the decision-making factors identified in Table-2. This phase has provided much deeper insights into the diverse information provided to the prospects and their contextual relevance to the decision-making factors.

The following example set shown in Table 4 is derived from the price attribute. This example set will establish the sequence of communications that will be sent to a prospect under-"price" profile. Tables 29, 30 and 31 in appendix – a will provide more contextually categorized pages that are extracted from the institutional website.

Table 4: Example set derived from the web and text mining phase for the context - Price

Link	URL
Cost by Major/Program	/admissions/financial-aid-tuition/tuition-and-fees/cost-by-major-program

Tuition and Fees	/graduate-students/graduate-admissions/tuition-and-fees
Tuition and Fees	/admissions/financial-aid-tuition/tuition-and-fees
Indirect Cost Policy	/assets/uploads/policies/01-73-00.pdf
Consumer Information	/about/consumer-information
Payment Methods Cost to Participate Simple Steps to Enroll Target …	/assets/uploads/resources/FACTS-bookmark-2015-2016.pdf
Undergraduate Student Fee	/assets/uploads/general/Undergrad-Tuition-Fees-2014-2015.pdf
Financial aid	/assets/uploads/resources/2015-16_Financial_Aid.pdf
Moving Expenses Reimbursement Guide	/assets/uploads/resources/Moving-Expenses-Reimbursement-Guide.pdf

4.3 Phase 2

As discussed in the data collection phase, the server logs contain unnecessary client-server communications that are to be processed in order to identify individual sessions. In the process of cleaning, only the date, time, client IP, client accessed URL's are extracted from the server logs. The client accessed URL's contain information about undergraduate and graduate programs the institution offers to prospects. This research has been limited to the graduate programs. Considering the fact that all these records are a combination of new versus returning and existing visitor information individual sessions are identified based on date, time, client IP and the URL's accessed.

The sessions are classified into three discrete phases named Pre-inquiry phase "Pre", Inquiry phase "Inq" and Post-inquiry phase "Post" based on the time, client IP address and URL. The inquiry phases are determined based on the inquiry page. Whenever prospect inquiries about a program he/she is redirected to a thank you, confirmation page. The confirmation page acts as a distinguishing point of inquiry and based on the IP classification, pages accessed the pre-inquiry; inquiry and post-inquiry phases are identified as shown in Figure 10. All the pre-inquiry, inquiry and post-inquiry pages accessed during the visit will act as individual sessions. As shown in table 5, from over a million records in raw server logs, 3590 records were identified as the number of inquiries from prospects who accessed around 14673 pages before making those inquiries and

26064 pages after making the inquiries. A total of 3590 individual sessions were extracted from the raw server logs.

Table 5: Inquiry, pre-inquiry and post-inquiry records identified from server logs

Year 2013	**Total**
Inquiry	3590
pre-inquiry	14673
post-inquiry	26064

Date	Time	Ip	URL	Phase
9/18/13	9:29:20	#	/mshi/health-informatics	pre
9/18/13	9:31:03	#	/gradoffice/mshi-confirm	inq
9/18/13	9:32:16	#	/mshi/	post
9/18/13	9:32:21	#	/gradoffice/grad-accreditation	post
9/18/13	9:33:11	#	/gradoffice/grad-tuition	post

Figure 10: Processed Server Logs

4.4 Phase 3

In this clustering and profiling phase, a prospective student classification is made based on the sessions identified in the pre-processing phase. This research study intends to capture the prospects' navigation patterns. In general, navigation patterns are best described as common browsing patterns among a group of visitors (Mustapha, Jalali, Bozorgniya, & Jalali, 2009). Visitors accessing specific information on a website tend to exhibit common interest up to a point during their time of navigation and the navigation patterns might exhibit those common interests or patterns. Different clustering techniques are tested to identify the technique that would be the best fit to this problem.

K-means clustering technique has been used to cluster the sessions identified in the pre-processing phase. The Euclidean distance measure is applied in Simple K-means clustering. Five different student profiles namely price, program, future employment, institutional image and environment are extracted from the literature. As the pre-determined number of clusters are extracted from K-means clustering, five clusters are extracted from this phase. However, the clusters obtained through this method are not uniformly distributed to form reliable clusters, so a different clustering technique “Expectation-Maximization Technique” was adapted to extract meaningful evenly distributed clusters.

The Expectation Maximization technique identifies the maximum likelihood estimates of the parameters and captures those overlapping interests. The output of the expectation maximization technique represents that with each iteration the log likelihood converges towards lower values up to a certain point and converge. This approach is performed in two steps: Expectation step and Maximization step. During the expectation step, the probability values for the clusters are computed and the maximization step will compute the likelihood of distribution parameters.

The IIS server log file consisted of different parameters such as date, time, server IP, URL, client IP and different server status messages. From all those parameters only date, time, client IP and client accessed URL are extracted in the pre-processing phase. Date and time played a crucial role in extracting individual sessions from the server logs.

This study intends to look for the prospect sessions with an inquiry. The pre-processed individual sessions consist of pre-inquiry, inquiry, and post-inquiry information. The pre-processed dataset comprises of five parameters namely date, time, client IP-address, URL accessed and the inquiry phase of the URL and although it has a limited contribution towards individual clusters date is a binding factor that completes individual sessions. After applying the clustering technique five different clusters are formed based on IP, URL, date, time and inquiry phase. The clustered instances have a common attribute URL based on which the inferences about the navigation patterns are to be made. The pre-processed dataset comprised of 3590 prospects who accessed about 175 web pages in about 44327 instances as shown in Table 5.

The clusters provided the frequently accessed web pages or the navigation patterns, but this study intends to extract the contextual sequence from the clustered instances. These contextual sequences are extracted by making use of the content profiles identified from the institutional website in the web and text mining phase and visualizing the clusters in a pivot table. The pivot table shown in Figure 11 is structured with clusters as columns and URL's as rows and having the count of IP's for each URL under a specific cluster as shown in the figure below.

	Count of ip	Column Labels					
4		cluster0	cluster1	cluster2	cluster3	cluster4	Grand Total
5	Row Labels						
6	/academics/academic-affairs/mset	6	15		1	8	30
7	/academics/academic-affairs/msis	19	29		18	15	81
8	/doctor-of-science/dscis-form	14	4		4	10	32
9	/doctor-of-science/dsis-msis-required-courses	80	68		45	50	243
10	/gradoffice		2				2
11	/gradoffice/	1535	849		523	722	3629
12	/gradoffice/business-analytics	89	92		46	79	306
13	/gradoffice/dscis-confirm			582	3		585
14	/gradoffice/dscis-form	2	1		4	3	10
15	/gradoffice/ethical-hacking	123	105		56	106	390
16	/gradoffice/grad-accreditation	170	112		94	118	494
17	/gradoffice/gradadmission		1				1
18	/gradoffice/grad-admission	635	583		627	619	2464
19	/gradoffice/grad-assistantship-deadlines	40	39		25	37	141
20	/gradoffice/grad-assistantships	79	67		50	72	268
21	/gradoffice/grad-contact	205	132		172	122	631
22	/gradoffice/grad-faculty	79	63		80	46	268
23	/gradoffice/grad-financial-aid	498	321		325	443	1587
24	/gradoffice/grad-forms	107	140		109	135	491
25	/gradoffice/grad-graduation-requirements	164	127		93	91	475
26	/gradoffice/grad-health-insurance	22	33		10	17	82
27	/gradoffice/grad-international	345	292		285	345	1267
28	/gradoffice/grad-news-resources	29	48		27	23	127
29	/gradoffice/grad-register	19	17		13	17	66

Figure 11: Clustering result sample

The contextual sequence is organized in a descending order based on the frequency of visits as shown in the below table 6. This will contribute to the mass-customized communication sequence.

Table 6: Content profiles prioritized based on the frequency of visits

Cluster 0	**instances classified 14673**
Program	6350
Price	5620
institutional image	1345
future employment	790
environment	568
Cluster 1	**instances classified 9232**

price	6082
institutional image	2092
program	927
future employment	67
environment	64
cluster 2	**instances classified 7250**
institutional image	4627
program	1410
price	785
future employment	295
environment	133
cluster 3	**instances classified 8642**
future employment	4670
program	1443
institutional image	1245
price	842
environment	442

The clusters are visualized in the form of a pivot table which provided more insights into the profiles. From the pivot view, most frequently accessed web pages are identified and sorted based on the frequency of visits. From the pivot views shown in Figure 11, four different sequences are extracted as shown in table 7. These four sequences derived that prospects in this cluster are more interested in the admission requirements, tuition fees, financial aid, international student requirements and assistantship information (note: The information is listed in descending order of popularity) based on the pages accessed. Figures 12, 13 and 14 provide a clear visualization of the IP, inquiry phase and URL distributions along with the clusters. The email communications in the mass-customized communication sequences are contextually prioritized based on the page frequencies in different clusters.

Table 7: Mass-customized sequences extracted from the clustering results

Sequence – 1	**Sequence - 2**	**Sequence -3**	**Sequence – 4**
Program	Price	institutional image	future employment
Price	institutional image	Program	Program
institutional image	Program	Price	institutional image
future employment	future employment	future employment	Price

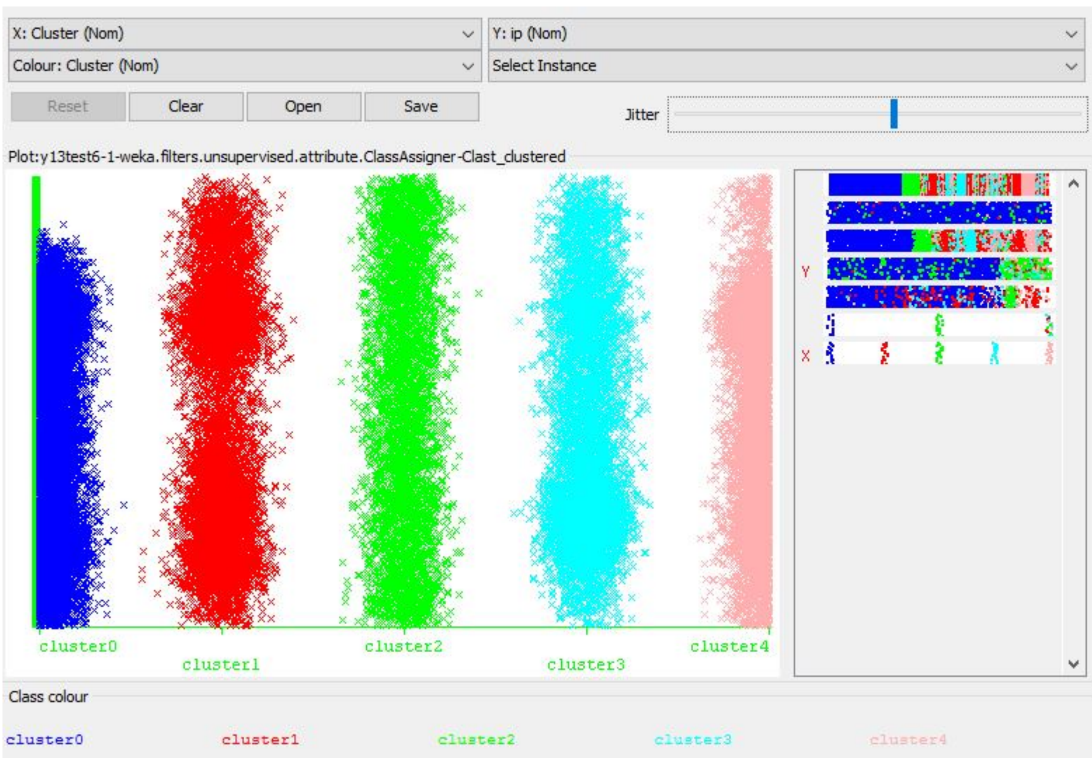

Figure 12: Clusters with respect to IP addresses

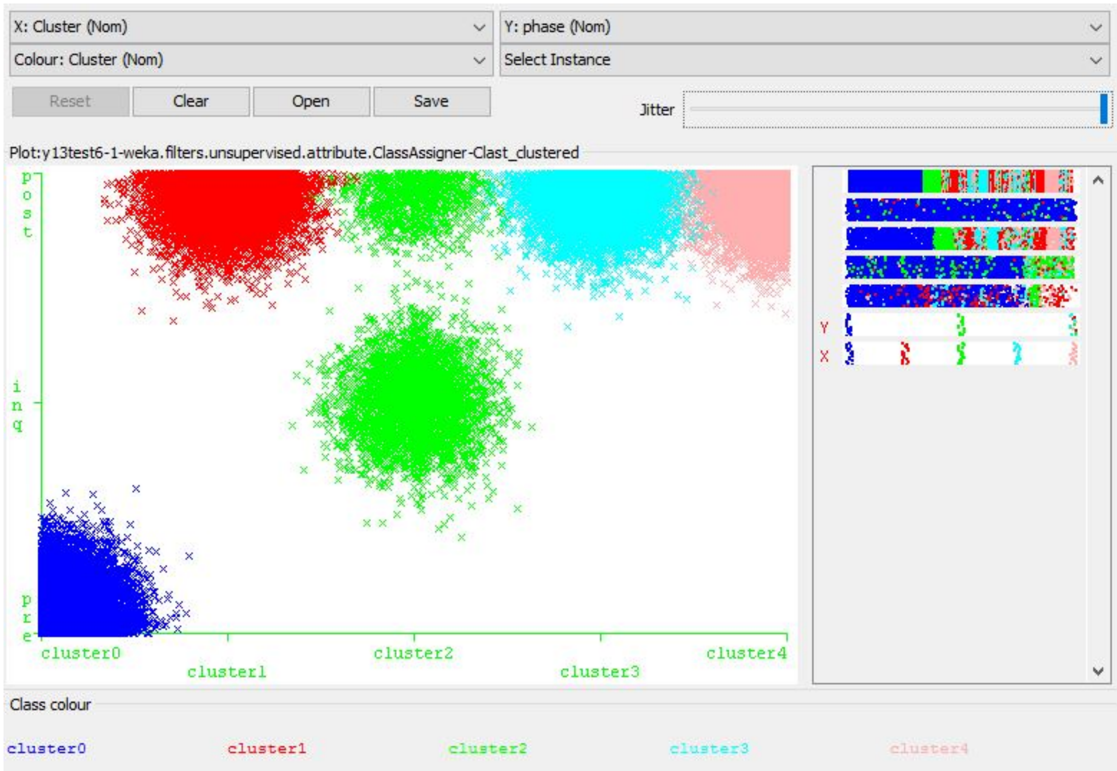

Figure 13: Clusters with respect to inquiry phase

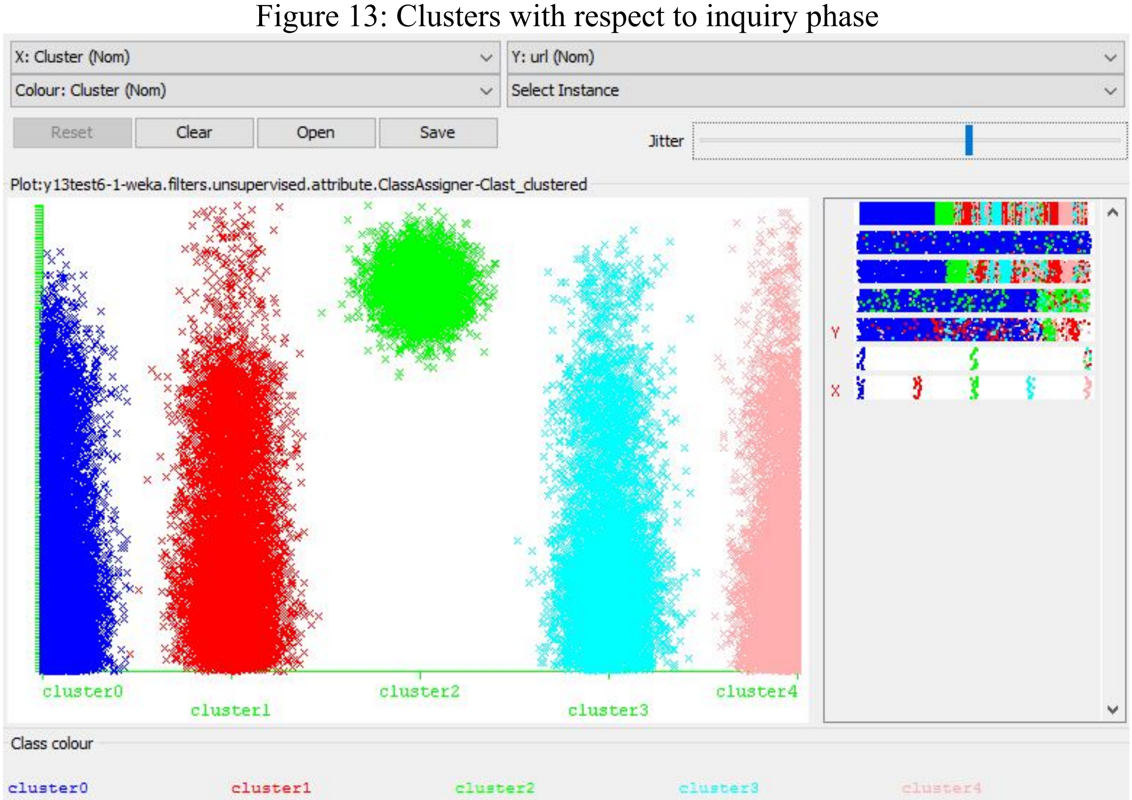

Figure 14: Clusters with respect to the URL's

4.5 Subject lines survey:

As stated in section 2.2.3.5, the number of emails people receive every day has increased exponentially. Individuals started prioritizing the emails for attention over others. Most people tend to spend a fraction of a second evaluating email subject fields before moving on to the next email. Extensive research has been conducted regarding the email surface features that draw the recipients attention to it (Wainer et al., 2011). Subject lines play an important role in capturing such attention and later incline him/her to perform an action on the email received. It is estimated that subject lines with 50 characters or less result in 12% higher open rates than those emails with longer subject lines (POP, 2015). To determine the relevance of subject lines with the prospective student profiles identified from the literature, a pilot study with the structured questionnaire was developed. The pre-structured questionnaire consisted of four questions representing the email subject lines and different contexts they might come under. The instrument was designed to allow the participant to select one context for each question. The sample size for the survey instrument was calculated using the following formula:

$$n = (t^2 * p(1 - p))/ m^2$$

Where:

n = required sample size

t = confidence level at 95% (standard value of 1.96)

p = estimated prevalence of the variable of interest (e.g. 20% or 0.2)

m = margin of error at 5% (standard value of 0.05)

Therefore: n = (3.84*1.6)/ (0.25) = 24.5

The questionnaire was posted on social media websites and the responses were collected from anonymous users. 25 responses are considered to be statistically significant for this study and the survey was conducted until 25 responses were reached.

The pilot survey was conducted with the following questionnaire to determine the contextual relevance with respect to the subject lines and the results were obtained in the following way: Out of 25 participants, 25 of them successfully completed the survey questionnaire. The responses were collected anonymously and are displayed in the below tables 8 and 9.

Table 8: Subject line survey questionnaire with responses

1	Know more about the program from our online catalog!	**Responses**
	Price	0%
	Program	100%
	Institutional Image	0%
	Future Employment	0%
2	Invest in yourself for your future and career advancement!	
	Price	0%
	Program	0%
	Institutional Image	0%
	Future Employment	100%
3	Dakota State University is Affordable!	
	Price	100%
	Program	0%
	Institutional Image	0%
	Future Employment	0%
4	What our students are saying about the image of DSU!	
	Price	0%
	Program	0%
	Institutional Image	100%
	Future Employment	0%

Table 9: Subject line survey results

Question No.	Intended context	Response
1	Program	100%
2	Future employment	100%
3	Price	100%
4	Institutional image	100%

4.6 Phase 4

Keeping the prospects engaged when they are in between the inquiry and application stages in the admission funnel is a crucial task. Phase 4 of the framework, the communication and relationship management phase tries to address the prospect engagement by making use of the mass-customized communication sequences that are identified in the clustering and profiling phase.

The existing prospective student communication system in the selected public university consists of email communications that are sent out in a pre-defined time frame. The time frame was set based on the date of inquiry and anticipated entry date into the program. There is a 7-day difference between one email communication to another. The existing communication system lacks the ability to provide customized information to the prospects that address their need

Field testing method will be applied to test the performance of mass-customized communications over the existing communication sequence. Prospect information for the months of May and June of the year 2016 were considered for the sample set. A total of 303 inquiries were made during that time frame. The control group and treatment group were selected through random sampling. The control group consists of 152 prospects and the remaining 151 prospects were categorized into four different treatments based on their browsing history. Out of 151 prospects, 40 prospects were categorized under treatment 1. Existing communication sequences were sent to the control group of 152 prospects and mass-customized communication sequences were sent to different treatments within the treatment group. Table 10 shown below provides more information on the communications the control and treatment groups would receive during the testing time frame.

Table 10: Existing sequence versus the mass-customized sequences

W	Mass Customized Seq. --> Existing Seq. ↓	**Seq1 –** (Program)	**Seq2 –** (Price)	**Seq3 –** (Ins. Image)	**Seq4 –** (Fut. Emp.)
1	Program	Program-	Price	Institutional Image	Future Employment
2	Institutional Image	Price-	Institutional Image	Program	Program
3	Price	Institutional Image	Program	Price	Institutional Image
4	Future Employment	Future Employment	Future Employment	Future Employment	Price

4.7 Evaluation of Framework

This study has evaluated the proposed framework through a controlled field experiment. The email communications were sent out to prospects in a control group as well as a treatment group. The treatment group consisted of four different treatments, each categorized by the contextual preference. The chi-square test was used to verify the statistical significance of the email open rates. The overall email open rate between the control and treatment groups was compared. Secondly, the email open rate of the control group and individual treatments was analyzed. Finally, the contextual difference of the emails opens in the control and first round email opens of each treatment in the treatment groups were analyzed.

The first round of email communications was sent out on July 25th of 2016 and the last round of email communications were sent on the 15th of August 2016. With an interval of 7 days between each email communication, the effect of a change in email communication sequence on the prospect engagement was measured in terms of the email open rate.

The average open rate of all the email communications was 21.12%, and the weekly open rates are shown in table 11 below.

Table 11: Weekly open rate for all email communications

	C	T1	T2	T3	T4
w1	17.10%	67.50%	42.85%	2.43%	*0%*
w2	9.80%	70%	45.70%	7.30%	*14.70%*

w3	21.71%	15%	17.14%	41.43%	*23.52%*
w4	12.50%	32.50%	28.50%	29.26%	*26.47%*
Average	15.28%	46.25%	33.55%	20.11%	16.17%

4.7.1 Evaluation of the control group and treatment group:

The existing sequence of email communications was sent out to the 152 prospects in the control group and mass-customized sequences of email communications were sent out to the 151 prospects within the treatment group. In the control group, from 608 emails sent out to all the prospects across four weeks, 77 emails were opened and 531 emails were not opened. In the treatment group, a total of 178 emails were opened and 426 emails were not opened. A 2×2 contingency table was formulated with the total number of emails opened for the emails was sent to both the groups.

The hypotheses as stated in section 3.5.1,

Null Hypothesis - H_0: The likelihood of a response from a prospect to the email communications remains same regardless of the change in the sequence of communications he/she receives from the institutions.

Let "x" be the response rate to the email communications and "mu" (μ) be the population of the control group.

$$H_0: \mu = X$$

Alternative Hypotheses:

- H_{a1}: The likelihood of a response from a prospect to the mass-customized email communications increases with the change in the sequence of communications he/she receives.

$$H_{a1}: \mu > X$$

- H_{a2}**:** The likelihood of a response from a prospect to the mass-customized email communications decreases with the change in the sequence of communications he/she receives.

$$H_{a2}: \mu < X$$

With a predetermined alpha level of significance α=0.05, a 2×2 contingency table with the number of emails opened from the control and treatment groups is shown in the below table 12.

Table 12: Contingency table for the control group and treatment group

	Opened	Not-Opened	Total
Control	77	531	608
Treatment	178	426	604
Total	252	957	1212

For a 2×2 contingency table the Chi-square statistic is calculated by the formula:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Where, f_o is the observed frequency and f_e is the expected frequency.

$$f_e = \frac{(N_r)(N_c)}{N}$$

Where N_r represents the total number of cases in the respective row, N_c represents the total number in the respective column, and N is the number in the full sample. The expected frequencies were calculated as shown below in table 13.

Table 13: Expected frequency table for control and treatment groups

	Opened	Not-Opened
Control	$f_e = 126.41$	$f_e = 480.07$
Treatment	$f_e = 125.58$	$f_e = 476.92$

The chi-square significance for control and treatment groups is calculated in the following way:

$$\text{Chi-square } \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= 52.02$$

The degree of freedom was calculated as df = (r-1) (c-1) = (2-1) (2-1) = 1, Where, r = the number of rows and c = the number of columns.

From the Chi-square distribution table, it is evident that the critical value with $df = 1$ and $\alpha = .05$ is 3.84. Because the calculated value exceeds this critical value, the difference is significant. The critical value for the obtained Chi-square of 52.02 at $df = 1$ is less than 0.0001 which means the p-value is below the predetermined level of significance and hence the null hypothesis H_0 is rejected.

The communications in the control group were opened by 77 prospects out of 608 prospects across all four weeks. The open rate of email communications in the control group was 12.66%. A total of 178 prospects out of 604 opened the email communications in the treatment group. The open rate of the email communications in the treatment group was 29.47%. With a change in the sequence of communications, there is a significant improvement in the email open rate of the treatment group to that of the control group. Hence, we accept the alternative hypothesis H_{a1}: The likelihood of a response from a prospect to the mass-customized email communication increases with a change in the sequence of communications he/she receives.

4.7.2 Evaluation of the control group and individual treatments:

The existing sequence of email communications was sent out to the 152 prospects in the control group and mass-customized sequences of email communications were sent out to the prospects in individual treatments.

In the control group, from a total of 608 emails sent out to all the prospects across four weeks, 77 emails were opened and 531 emails were not opened. Treatment 1 consisted of 40 prospects with a total of 160 emails sent to the prospects in a four-week time frame. In treatment 1, 74 emails were opened and 86 emails were not opened. Treatment 2 consisted of 35 prospects with a total of 140 emails sent to the prospects in a four-week time frame. In treatment 2, 47 emails were opened and 93 emails were not opened. Treatment 3 consisted of 41 prospects with a total of 164 emails sent to prospects in a four-week time frame. In treatment 3, 35 emails were opened and 129 emails were not opened. Treatment 4 consisted of 30 prospects with a total of 120 emails sent to prospects in a four-week time frame. In treatment 4, 22 emails were opened and 114 emails were not opened.

A 2×2 contingency table depicting the total number of emails opened and not opened was provided for the control group and individual treatments that are shown in the tables 14, 16, 18 and 20 respectively. The expected frequencies for the control group and individual treatments were calculated as shown in table 15, 17, 19 and 21 respectively.

Table 14: Contingency table for control group and treatment 1

	Opened	Not-Opened	Total
Control	77	531	608
Treatment 1	74	86	160
Total	151	617	768

Table 15: Expected frequency table for control group and treatment 1

	Opened	Not-Opened
Control	$f_e = 119.54$	$f_e = 488.45$
Treatment 1	$f_e = 31.45$	$f_e = 128.54$

Table 16: Contingency table for control group and treatment 2

	Opened	Not-Opened	Total
Control	77	531	608
Treatment 2	47	93	140
Total	124	624	748

Table 17: Expected frequency table for control group and treatment 2

	Opened	Not-Opened
Control	$f_e = 100.79$	$f_e = 507.20$
Treatment 2	$f_e = 23.20$	$f_e = 116.79$

Table 18: Contingency table for control group and treatment 3

	Opened	Not-Opened	Total
Control	77	531	608
Treatment 3	35	129	164
Total	112	660	772

Table 19: Expected frequency table for control group and treatment 3

	Opened	Not-Opened
Control	$f_e = 88.20$	$f_e = 510.79$
Treatment 3	$f_e = 23.79$	$f_e = 140.20$

Table 20: Contingency table for control group and treatment 4

	Opened	Not-Opened	Total
Control	77	531	608
Treatment 4	22	114	136
Total	99	645	744

Table 21: Expected frequency table for control group and treatment 4

	Opened	Not-Opened
Control	$f_e = 80.90$	$f_e = 527.09$
Treatment 4	$f_e = 18.09$	$f_e = 117.90$

The chi-square significance for control group and treatments were calculated in the following way:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Table 22: Chi-square values for individual treatments

Treatments	**Chi-Square Values**
Treatment 1	90.46
Treatment 2	35.98
Treatment 3	8.39
Treatment 4	1.19

The degree of freedom was calculated as df = (r-1) (c-1) = (2-1) (2-1) = 1, Where, r = the number of rows and c = the number of columns

From the Chi-square distribution table, it is evident that the critical value with *df* = 1 and α = .05 is 3.84. Because the calculated value for the control group and treatment 1 exceeds this critical value, the difference is significant. The critical value for the obtained Chi-square of 90.46 at *df = 1* is less than 0.0001 which means the *p*-value is below the predetermined level of significance and hence the null hypothesis H_0 is rejected.

The open rate of email communications in the control group was 12.66%. Whereas, the open rates of the email communications in the treatment 1, 2 and 3 were 46.25%, 33.57%, and 21.34% respectively. This clearly indicated that, with a change in the sequence of communications, there is a significant improvement in the email open rate of the treatments 1, 2 and 3 to that of the control group. Hence, we accept the alternative hypothesis H_{a1}: The likelihood of a response from a prospect to the mass-customized email communication increases with a change in the sequence of communications he/she receives.

The chi-square significance for control group and treatment 4 was $\chi^2 = 1.19$, From the Chi-square distribution table, it is evident that the critical value with *df* = 1 and α = .05 is 3.84. Because the calculated value is less than this critical value, the difference is not significant. The critical value for the obtained Chi-square of 1.19 at *df = 1* is greater than 0.05 which means the *p*-value is above the predetermined level of significance for treatment 4 and hence the study fails to reject null hypothesis H_0. There is no statistically significant difference in terms of the email open rates observed between the control group and treatment 4.

The average open rate of the emails in control group was 15.28%, whereas the four treatments in the treatment group received 46.25%, 33.55%, 20.11% and 16.17% respectively. The average open rates for all the treatments in the treatment group surpassed the control group open rate, which clearly indicates the success of framework in providing better engagement. The prospects in each treatment were contextually categorized and the following section structured to provide a contextual evaluation of each email in control group to that of the treatments in the treatment group.

4.7.3 Contextual Evaluation

In both control group and the individual treatment groups different context oriented emails were sent out the prospects. The communication sequences in the treatment group were mass-customized and the prospects were contextually categorized based on their behavior. Although the emails in control group and treatments were sent at the same time, the context of the emails varied within the sequence from the control group to the individual treatments in the treatment group. This contextual evaluation was an attempt to understand whether the first round of emails in the individual treatments would make a significant difference.

The hypotheses as stated in section 3.5.2,

Null Hypothesis - H_0: The likelihood of a response from a prospect to the email communications remains same regardless of the change in the sequence of communications he/she receives that matches the contextual preference of the prospect.

Let "x" be the response rate to the email communications and "mu" (μ) be the population of the control group.

$$H_0: \mu = X$$

Alternative Hypotheses:

- H_{a1}: The likelihood of a response from a prospect to the mass-customized email communications increases with the change in the sequence of communications he/she receives that matches the contextual preference of the prospect.

$$H_{a1}: \mu > X$$

- H_{a2}**:** The likelihood of a response from a prospect to the mass-customized email communications decreases with the change in the sequence of communications he/she receives that matches the contextual preference of the prospect.

$$H_{a2}: \mu < X$$

The email communication, program, was sent on week 1 in both control group and treatment 1 in the treatment group. A total of 26 prospects out of 152 opened the program email in the control group, and a total of 27 prospects out of 40 opened the program email in treatment group 1. In the control group, the email communication, price, was sent on week 3, whereas in the treatment 2 the email communication price was sent on week 1. A total of 18 prospects out of 152 opened the price email communication in the control group and a total of 15 prospects from 35 opened the price email communication in the treatment 2. In the control group, the email communication, institutional image, was sent in week 2, whereas in the treatment 3 the email communication, institutional image, was sent in week 1. A total of 15 prospects out of 152 opened the institutional image email communication in the control group and only one prospect from 41 opened the institutional image email communication in the treatment 3. In the control group, the email communication, future employment, was sent on week 4, whereas in the treatment 4 it was sent in week 1. A total of 19 prospects out of 152 opened the email in the control group and 0 prospects out of 34 opened it in treatment 4.

A 2×2 contingency table was formulated for each context as shown in tables 22, 24, 26, and 28 respectively. The contingency tables have provided the total number of emails opened when the email was sent to control group and treatments.

Table 23: Contingency table for the context program

Context – Program	Opened	Not-Opened	Total
Control	26	126	152
Treatment	27	13	40
Total	53	139	192

With a predetermined alpha level of significance α=0.05 and a 2×2 contingency table the Chi-square statistic is calculated by the formula:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Where, f_o is the observed frequency and f_e is the expected frequency.

$$f_e = \frac{(N_r)(N_c)}{N}$$

Where N_r represents the total number of cases in the respective row, N_c represents the total number in the respective column, and N is the number in the full sample. The expected frequencies for the context program were calculated as shown in tables 23, 25, 27 and 29 respectively.

Table 24: Expected frequency of the context program

	Opened	Not-Opened
Control	$f_e = 41.95$	$f_e = 110.04$
Treatment	$f_e = 11.04$	$f_e = 28.95$

Table 25: Contingency table for the context price

Context - Price	Opened	Not-Opened	Total
Control	18	134	152
Treatment	15	20	35
Total	33	154	187

Table 26: Expected frequency of the context price

	Opened	Not-Opened
Control	$f_e = 26.82$	$f_e = 125.17$
Treatment	$f_e = 6.17$	$f_e = 28.82$

Table 27: Contingency table for the context - institutional image

Context – Institutional Image	Opened	Not-Opened	Total
Control	15	137	152
Treatment	1	40	41
Total	16	177	193

Table 28: Expected frequency of the context institutional image

	Opened	Not-Opened
Control	$f_e = 12.60$	$f_e = 139.39$
Treatment	$f_e = 3.39$	$f_e = 37.60$

Table 29: Contingency table for the context - future employment

Context – Future Employment	Opened	Not-Opened	Total
Control	19	131	152
Treatment	0	34	34
Total	19	165	186

Table 30: Expected frequency of the context future employment

	Opened	Not-Opened
Control	$f_e = 15.52$	$f_e = 134.83$
Treatment	$f_e = 3.47$	$f_e = 30.16$

The chi-square significance for the context program is calculated in the following way, and the chi-squares for individual contexts was provided in table 30:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Table 31: Chi-square values for individual contexts

Context	Chi-Square
Program	40.25
Price	18.84
Institutional Image	2.32
Future Employment	4.84

The degree of freedom was calculated as df = (r-1) (c-1) = (2-1) (2-1) = 1*,* Where, r = the number of rows and c = the number of columns.

From the Chi-square distribution table, it is evident that the critical value with *df* = 1 at α = .05 is 3.84.

For the contexts' program and price, the chi-square values were calculated as 40.25 and 18.24 respectively. The calculated values have exceeded the critical value and the difference is significant. The critical value for the obtained chi-squares at *df = 1* is less than 0.0001, which means the *p*-value is below the predetermined level of significance and hence the null hypothesis H_0 was rejected for both contexts. 17.10% of the control group prospects' and 67.50% of the treatment 1 prospects opened program email. 11.8% of the control group prospects and 60% of the treatment 2 prospects opened price email. From the obtained open rates, it is evident that, with a change in the sequence of communication that matches the contextual preference of the prospect, there is a significant improvement in the email open rate of the treatment group to that of the control group. Hence, we accept the alternative hypothesis H_{a1}: The likelihood of a response from a prospect to the mass-customized email communication increases with a change in the sequence of communications he/she receives that matches the contextual preference of the prospect.

For the context institutional image, the calculated chi-square value is less than the critical value, and hence there is no significant difference. The critical value for the obtained chi-square of 2.32 at *df = 1* is 0.127, which means the *p*-value is above the predetermined level of significance and hence the study fails to reject the null hypothesis H_0. 9.8% and the control group prospects and 2.5% of the prospects in the treatment 3 opened institutional image email. Hence we accept the alternative hypothesis H_{a2}: The likelihood of a response from a prospect to the mass-customized email communication decreased with a change in the sequence of communications he/she receives that matches the contextual preference of the prospect.

For the context future employment, the calculated chi-square value exceeds the critical value, the difference is significant. The critical value for the obtained chi-square of 4.84 at *df = 1* is 0.0278 which means the *p*-value is below the predetermined level of significance and hence the null hypothesis H_0 was rejected. 12.5%. of the prospects in control group and 0% of the prospects in treatment 4 opened the future employment email. This states that, with a change in the sequence of communication that matches the contextual preference of the prospect, there is a significant decrease in the email open rate of the treatment group to that of the control group. Hence, we accept the alternative hypothesis H_{a2}: The likelihood of a response from a prospect to the mass-customized email communication decreased with a change in the sequence of communications he/she receives that matches the contextual preference of the prospect.

This research statistically validated the hypothesis that states that the sequence of mass-customized communications that are sent to prospects have a significant impact on the prospective student response rate as well as the prospect engagement with the institution. The treatment group, consisting of mass-customized communication sequences, had over 16% better open rates than the control group. Individual treatments also surpassed the control group open rates as shown in figure 15 in appendix b. With the contextual evaluation of the communications in control group and treatment group, it became evident that price and program received more open rates than the remaining contexts institutional image and future employment.

Chapter 5 Discussion and Future Research

5.1 Contributions

Existing literature had traces of research that made use of historical prospective student information to predict the student enrollments into the institutions, but no specific evidence of engaging prospective students through behavioral mining concepts was found. Predictions that were made by depending on the prospect inquiries do not provide accurate results when the institutions receive partial inquiries. No evidence from existing literature was found that made use of prospect navigational behavior from institutional server logs. This framework is designed to extract prospect navigational behavior from historical server logs and help the institutions in optimizing their engagement strategies.

This research is based on the underpinnings of prospect decision-making variables identified from existing higher education marketing literature. These decision-making variables were then translated into web content profiles and prospect profiles based on their contextual relevance.

This research contributes to the higher education marketing body of knowledge by providing a unique analytics-based solution that extracts web content profiles as well as the prospect profiles and contributed to the institutional marketing efforts to better engage the prospects in the higher-education sales cycle. This study has extended the use of analytics techniques in a new dimension, i.e. higher education marketing. This study has identified prospect's pre-inquiry, inquiry and post-inquiry navigational information as crucial elements in determining the communication strategy. This study has introduced mass-customized communications to replace the existing mass communication sequences

The framework was successfully tested in a higher education institution in Midwest. Additionally, this research evaluated the prospective student response rates to the customized sequence of communications with statistically significant results. The results of the study can be used as a baseline for other higher education institutions to understand their prospect needs and this framework can be adopted to profile their prospects and enhance their engagement strategies.

The framework comprised of four distinct phases, the web and text mining phase, pre-processing phase, clustering and profiling phase and the communication and relationship management phase. The web and text mining phase extracted the web page content profiles from contextually related web pages through content profiling. The pre-processing phase extracted individual sessions with pre-inquiry, inquiry and post-inquiry data from the institutional server logs. These sessions were then clustered in the clustering and profiling phase. The clusters were compared with the web page content profiles and the pages were categorized based on the frequency of visits. These categories represent mass-customized communication sequences. Higher education institutions can use this model to understand their prospect behavior and profile their prospect needs.

The framework was tested in a higher education institution in Midwest, United States. This research has successfully evaluated the framework by extracting the traces of prospective student navigational information hidden in institutional server logs. The results indicated the success of the framework when tested in the field experiment.

In summary, the proposed framework helps higher-education institutions optimize their prospect engagement strategy through the mass-customized communication sequences. Some of the special features of the framework include generalizability, user-friendliness, and reliability. This framework may be implemented in any higher educational setting that has an established website with traffic.

5.2 Limitations

There is a small probability that prospects might make inquiries from different computers about different programs, which would be recorded as multiple inquiries. There is also a chance that some of the prospects might be using email services with a preview window which will not download the whole message. Such interactions would not be recorded in the server logs and would not provide the actual number of emails opened. While the framework is flexible and able to provide the recommendations, it needs to be managed by someone knowledgeable enough to make necessary changes within the program to extract individual sessions from the data, cluster them and interpret the results in a manner similar to that performed in Chapter 4. Subject lines played an important role on the email open rates and the results of the subject line survey might not be replicable. This study was tested on graduate prospects and the behavior and profiles might vary for undergraduate prospects.

5.3 Future research and conclusion

In the future, this research can be applied to profile undergraduate prospects and understand their behavior. The framework can also be incorporated into the institutional setting by making use of expert systems and thereby automate different phases like pre-processing, clustering and mass-customizing the communication sequences.

This study has provided a unique and effective solution to extract profile based mass-customized communication sequences aimed to provide better engagement to the higher education prospects. This study has successfully analyzed the specific impact of mass-customized communications on the email open rates and also observed the influence of context over the email open rates. The field experiment has clearly demonstrated the impact of the framework on understanding prospect needs. This research argued that the application of analytic models in higher education institutions was limited to reporting purposes and provided a constructive solution to the existing prospect engagement problem. This research has also successfully tested the framework and demonstrated that the framework is capable enough to bring a significant difference in the prospect engagement within the institutional setting.

Appendices

Appendix A

Table 32: Program-related pages extracted from web-content-mining

Program Pages
/graduate-students/dsccs/timelines-and-requirements
/graduate-students/dscis/timelines-and-requirements
/graduate-students/mba/timelines-and-requirements
/graduate-students/msa/timelines-and-requirements
/graduate-students/msacs/timelines-and-requirements
/graduate-students/mset/timelines-and-requirements
/graduate-students/mshi/timelines-and-requirements
/graduate-students/msia/timelines-and-requirements
/graduate-students/msis/timelines-and-requirements
/graduate-students/banking-security-grad-certificate/course-rotation
/graduate-students/dscis/course-rotation
/graduate-students/information-technology-grad-certificate/course-rotation
/graduate-students/msa/course-rotation
/graduate-students/msacs/course-rotation
/graduate-students/mset/course-rotation
/graduate-students/mshi/course-rotation
/graduate-students/msia/course-rotation
/graduate-students/msis/course-rotation

Table 33: Institutional image pages extracted from web-content-mining

Institutional Image Pages
/about-dsu
/about-dsu/why-dsu
/about-dsu/campus-tour
/about-dsu/maps-and-directions
/about-dsu/accreditation
/about-dsu/directory
/about-dsu/directory/office-directory
/about-dsu/foundation
/about-dsu/foundation/give-to-dsu
/about-dsu/foundation/board-of-trustees
/about-dsu/foundation/foundation-staff
/about-dsu/consumer-information
/about-dsu/public-relations-and-marketing
/about-dsu/campus-construction
/graduate-students/dsccs/faculty

/graduate-students/dscis/faculty
/graduate-students/mba/faculty
/graduate-students/msa/faculty
/graduate-students/msacs/faculty
/graduate-students/mset/faculty
/graduate-students/mshi/faculty
/graduate-students/msia/faculty
/graduate-students/msis/faculty
/alumni-and-friends
/alumni-and-friends/give-a-gift
/alumni-and-friends/update-your-information
/alumni-and-friends/dsu-license-plates
/alumni-and-friends/alumni-magazine
/alumni-and-friends/alumni-testimonials
/alumni-and-friends/official-transcripts
/alumni-and-friends/alumni-board

Table 34: Future employment pages extracted from web-content-mining

Future Employment Pages
/student-life/career-services
/student-life/career-services/dsu-job-link
/student-life/career-services/outcome-survey
/student-life/career-services/job-fairs
/student-life/career-services/for-employers
/student-life/career-services/meet-our-staff

Appendix B

Sequence	Control Group (152)				Treatment Group-1: Program (40)				Treatment Group-2: Price (35)				Treatment Group-3-IImage (41)				Treatment Group-4: F-emp. (34)			
Week	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4
Context	program	iimage	price	femp	program	price	iimage	femp	price	iimage	program	femp	iimage	program	price	femp	femp	program	iimage	price
1	26	0	0	0	27	0	0	0	15	0	0	0	1	0	0	0	0	0	0	0
2	0	15	0	0	0	28	0	0	0	16	0	0	0	5	0	0	0	5	0	0
3	0	0	18	0	0	0	6	0	0	0	6	0	0	0	17	0	0	0	8	0
4	0	0	0	19	0	0	0	13	0	0	0	10	0	0	0	12	0	0	0	9
Total Opens	26	15	18	19	27	28	6	13	15	16	6	10	1	3	17	12	0	5	8	9
Open Rates	17.10%	9.80%	21.71%	12.50%	67.50%	70%	15%	32.50%	42.85%	45.70%	17.14%	28.50%	2.43%	7.30%	41.43%	29.26%	0%	14.70%	23.52%	26.47%

Figure 15: Overview of control and treatment group individual email opens and total open rates

Appendix C

The table below elucidates different knowledge factors acquired from the existing marketing literature that contains all the key performance indicators.

Note: KA – Knowledge Acquisition, KC – Knowledge Creation and KM – Knowledge Management

Table 35: Additional literature reviewed while framing the research problem

Author	Objective	Knowledge factors	Knowledge
(Adar, Weld, Bershad, & Gribble, 2007), (Punera & Merugu, 2010) (Raj, Dey, & Gaonkar, 2011)	Analyze World Wide Web data and identified behavioral patterns to understand predictive power.	Visitor/ User Behavior	KA, KC
(Apte et al., 2003; Apte, Natarajan, Pednault, & Tipu, 2002)	Probabilistic modeling for insurance risk management and text mining.	Scalability and Reliability	KA, KC
(Ashkan, Clarke, Agichetein, & Guo, 2009) (Regelson & Fain, 2006)	Predictive model to estimate the ad click through rate through query intent analysis	Click Through Rates	KA, KC
(Awad, Khan, & Thuraisingham, 2008)	Predict the WWW surfing by using multiple evidence combination	Latency (time delay), personalization	KA, KC
(Duggan & Payne, 2008)	A predictive model to predict user domain knowledge from search behaviors	Domain Knowledge	KA, KC
(Gruhl, Guha, Kumar, Novak, & Tomkins, 2005)	Correlated postings in blogs, media, and web to draw conclusions on predictive power	Queries	KC, KA
(Huffman & Hochster, 2007)	A predictive approach to approximate the visitor satisfaction based on query relevance	Query relevance	KC, KA
(Oliner, Ganapathi, & Xu, 2012)	Use of log analysis for making better predictions	Visitor Log	KC, KA, KM
(Introna & Nissenbaum, 2000)	Explain the role of indexing in achieving search engine recognition	Indexing	KC, KA, KM
(Sculley, Malkin, Basu, & Bayardo, 2009)	Proposed a predictive model to predict the bounce rates in sponsored search	Bounce Rates	KC, KA

(Wen, Nie, & Zhang, 2002)	Applying the knowledge of query clustering using user logs for identifying better method	Query clustering, keyword clustering	KC, KA, KM
(White, Dumais, & Teevan, 2009), (X. Zhang, Cole, Street, & Belkin, 2011)	Characterize the influence of domain expertise on web search behavior	Domain knowledge and user behavior	KC, KA
(Ghose & Yang, 2008)	Analyzed the firm's behavior for predicting the E-marketing performance	Behavior	KC, KA
(Zhu et al., 2009)	Analyze historical click-through data to optimize the search engine revenue	Click-Through Data	KC, KA
(Yang, Va, & Zhang, 2001)	Applying the web log mining for higher precision of advertising	Patterns from historical data	KC, KA

References

Aarinen, A. (2012). *University Websites As Facilitators Of International Student Decision Making*.

Adar, E., Weld, D. S., Bershad, B. N., & Gribble, S. D. (2007). Why We Search : Visualizing and Predicting User Behavior. *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*.

Adomavicius, G., & Alexander, T. (2001). Using Data Mining Methods to Build Customer Profiles. *IEEE Computer*, 74–82.

Ali, D. M. M. (2013). Role of Data Mining In Education Sector. *International Journal of Computer Science and Mobile Computing*, *2*(4), 374–383.

Appling, J. G., & College, C. (2012). We ' re Going Google ! Making the Most of Marketing. In *In Proceedings of the ACM SIGUCCS 40th annual conference on Special interest group on university and college computing services (SIGUCCS '12)* (pp. 223–226).

Apte, C. V., Hong, S. J., Natarajan, R., Pednault, E. P. D., Tipu, F. a., & Weiss, S. M. (2003). Data-intensive analytics for predictive modeling. *IBM Journal of Research and Development*, *47*(1), 17–23. http://doi.org/10.1147/rd.471.0017

Apte, C. V., Natarajan, R., Pednault, E. P. D., & Tipu, F. a. (2002). A probabilistic estimation framework for predictive modeling analytics. *IBM Systems Journal*, *41*(3), 438–448. http://doi.org/10.1147/sj.413.0438

Ash, T. (2008). *Landing Page Optimization: The definitive guide to testing and tuning for conversions*. Book, Wiley Publishers.

Ashkan, A., Clarke, C. L. A., Agichetein, E., & Guo, Q. (2009). Estimating Ad Clickthrough Rate through Query Intent Analysis. *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT '09)*, *1*.

Awad, M., Khan, L., & Thuraisingham, B. (2008). Predicting WWW surfing using multiple evidence combination. *The VLDB Journal*, *17*(3), 401–417. http://doi.org/10.1007/s00778-006-0014-1

Bamshad Mobasher, Namit Jain, Eui-Hong Han, J. S. (1997). Web Mining: Pattern discovery from World Wide Web transactions.

Bhate, D. V, & Pasha, M. Y. (2014). Analyzing Target Customer Behavior Using Data Mining

Techniques for E- Commerce Data. *International Journal of Innovative Research in Computer Science & Technology*, *2*(1), 16–19.

Bichsel, J. (2012). *Analytics in Higher Education: Benefits, Barriers, Progress, and Recommendations*.

Catherine Bounsaythip, E. R.-R. (2001). *Overview of Data Mining for Customer Behavior Modeling*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.3279

Chen, M.-S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 866–883.

Corley, J. K., Jourdan, Z., & Ingram, W. R. (2013). Internet marketing: a content analysis of the research. *Electronic Markets*, *23*(3), 177–204. http://doi.org/10.1007/s12525-012-0118-y

Crossley, M., Kings, N. J., & Scott, J. R. (2003). Profiles — Analysis and Behaviour. *BT Technology Journal*, *21*(1), 56–66. http://doi.org/10.1023/A:1022404310934

Desjardins, S. L. (2002). An Analytic Strategy To Assist Institutional Recruitment And Marketing Efforts. *Research in Higher Education*, *43*(5), 531–553.

Duggan, G. B., & Payne, S. J. (2008). Knowledge in the Head and on the Web : Using Topic Expertise to Aid Search. *ACM*, 39–48.

Dupaul, S. (2010). *Understanding The Behaviors of Stealth Applicants In The College Search Process*.

Filip, A. (2012). Marketing theory applicability in higher education. *Procedia - Social and Behavioral Sciences*, *46*, 912 – 916. Retrieved from http://ac.els-cdn.com/S1877042812013523/1-s2.0-S1877042812013523-main.pdf?_tid=d1fd8e2a-d52e-11e5-8a18-00000aacb362&acdnat=1455683457_298c13f0aa1c6bd6c78be0b6e7c5a583

Fischer, R., Wittern, E., Schneider, J., & Tai, S. (2012). Pattern-based process discovery. In *Service-Oriented Computing and Applications (SOCA), 2012 5th IEEE International Conference on* (pp. 1–8). http://doi.org/10.1109/SOCA.2012.6449442

Ghose, A., & Yang, S. (2008). Analyzing Search Engine Advertising : Firm Behavior and Cross-Selling in Electronic Markets. *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, 219–226. Retrieved from http://doi.acm.org/10.1145/1367497.1367528

Gilroy, B. M. (2009). Migrates to YouTube and Social Networks, *19*, 12–14.

Goenner, C. F., & Pauls, K. (2006). A Predictive Model of Inquiry to Enrollment. *Research in Higher Education*, *47*(8), 935–956. http://doi.org/10.1007/s11162-006-9021-8

Gomes, L., & Murphy, J. (2003). An exploratory study of marketing international education online. *International Journal of Educational Management*, *17*(3), 116–125. http://doi.org/10.1108/09513540310467787

Grappone, J., & Couzin, G. (2006). *Search Engine Optimization: An hour A day*. Book, Wiley .

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05)* (pp. 78–87). ACM Press. http://doi.org/10.1145/1081870.1081883

Hemsley-Brown, J., & Oplatka, I. (2006). Universities in a Competitive Global Marketplace: A Systematic Review of The Literature on Higher Education Marketing. *International Journal of Public Sector Management*, *19*(4), 316–338. http://doi.org/10.1108/09513550610669176

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems. *MIS Quarterly*, *28*(1), 75–105.

Hossler, D., & Gallagher, K. S. (1987). Studying student college choice: A three-phase model and the implications for policymakers. *College and University*, *62*(3), 207–221.

Huang, J. (2003). Customer Choice between Electronic and Traditional Markets : An Economic Analysis. *Proceedings of the 36th Hawaii International Conference on System Sciences*. Retrieved from http://www.hicss.hawaii.edu/HICSS36/HICSSpapers/INMIW04.pdf

Huffman, S. B., & Hochster, M. (2007). How Well does Result Relevance Predict Session Satisfaction ? *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*, 567–573.

Husin, H. S. (2013). News recommendation based on Web usage and Web content mining. In *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)* (pp. 326–329). Ieee. http://doi.org/10.1109/ICDEW.2013.6547475

Introna, L., & Nissenbaum, H. (2000). Defining the Web: the politics of search engines. *Computer*, *33*(1), 54–62. http://doi.org/10.1109/2.816269

Ivy, J., & Naude, P. (2004). Succeeding in the MBA Marketplace: Identifying the Underlying Factors. *Journal of Higher Education Policy and Management*, *26*(3), 401–417.

Jones, K. B. (2008). *Search Engine Optimization* (Vol. Wiley Publ). Book.

Kathuria, A., Jansen, B. J., Hafernik, C., & Spink, A. (2010). Classifying the user intent of web queries using k-means clustering. *Internet Research*, *20*(5), 563–581. http://doi.org/10.1108/10662241011084112

Kelly, S. (2003). Mining data to discover customer segments. *Interactive Marketing*, *4*(3), 235–242.

Kiang, M. Y., Raghu, T. ., & Shang, K. H.-M. (2000). Marketing on the Internet — who can benefit from an online marketing approach? *Decision Support Systems*, *27*(4), 383–393. http://doi.org/10.1016/S0167-9236(99)00062-7

Kosala, R., & Blockeel, H. (2000). Web Mining Research: A Survey. *ACM SIGKDD Explorations Newsletter*, *2*(1), 1–15. http://doi.org/10.1145/360402.360406

Kotler, P., & Armstrong, G. (2006). *Principles of Marketing*.

Labus, M., & Stone, M. (2010). The CRM behaviour theory – Managing corporate customer relationships in service industries. *Journal of Database Marketing & Customer Strategy Management*, *17*(3–4), 155–173. http://doi.org/10.1057/dbm.2010.17

Liu, H., & Keselj, V. (2007). Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. *Data & Knowledge Engineering*, *61*(2), 304–330. http://doi.org/10.1016/j.datak.2006.06.001

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*, 251–266.

María Cubillo, J., Sánchez, J., & Cerviño, J. (2006). International students' decision- making process. *International Journal of Educational Management*, *20*(2), 101–115. http://doi.org/10.1108/09513540610646091

Markov, Z., & Larose, D. T. (2007). *Data Mining The Web - Uncovering Patterns in Web Content, Structure, and Usage*. John Wiley & Sons, Inc.

McCoy, J. C. (2011). A Comparison of Internet Marketing Methods Utilized by Higher Education Institutions.

Michael, B. P. (1990). *College Choice: Understanding Student Enrollment Behavior ASHE - ERIC Higher Education Report No.6*.

Moogan, Y. J. (2011). Can a Higher Education Institution's Marketing Strategy Improve the Student-Institution Match? *International Journal of Educational Management*, *25*(6), 570–589. http://doi.org/10.1108/09513541111159068

Moogan, Y. J., Baron, S., & Harris, K. (1999). Decision-Making Behaviour of Potential Higher Education Students. *Higher Education Quarterly*, *53*(3), 211–228. http://doi.org/10.1111/1468-2273.00127

Moran, M., & Hunt, B. (2009). *Search Engine Marketing, Inc. Driving search traffic to your company's website* (Second edi). Book, IBM.

Morris, C. G. (2009). *Prospective Student Interaction and Engagement with the College : Linking Timeliness to Enrollment Likelihood*.

Mustapha, N., Jalali, M., Bozorgniya, A., & Jalali, M. (2009). Navigation Patterns Mining Approach based on Expectation Maximization Algorithm. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *3*(2), 422–426.

Naidoo, V., & Connect, E. (2011). Marketing strategy implementation in higher education : A mixed approach for model development and testing, *27*(October), 1117–1141.

Nicolescu, L. (2009). Applying Marketing to Higher Education: Scope and Limits. *Management & Marketing*, *4*(2), 35–44.

Nicoletti, M., Schiaffino, S., & Godoy, D. (2013). Mining interests for user profiling in electronic conversations. *Expert Systems with Applications*, *40*(2), 638–645. http://doi.org/10.1016/j.eswa.2012.07.075

Noel-Levitz. (2012). *2012 Marketing and Student Recruitment Practices for Master 's-Level Graduate Programs*.

Oinas-kukkonen, H., & Harjumaa, M. (2009). Persuasive Systems Design: Key Issues , Process Model , and System Features. *Communications of the Association for Information Systems*, *24*(1), 1–18.

Oliner, A., Ganapathi, A., & Xu, W. (2012). Advances and challenges in log analysis. *Communications of the ACM 55*, *55*(2), 55–61. http://doi.org/10.1145/2076450.2076466

Oplatka, I., & Hemsley-Brown, J. (2004). The Research on School Marketing: Current Issues and Future Directions. *Journal of Educational Administration*, *42*(3), 375–400. http://doi.org/10.1108/09578230410534685

Peffers, K., Tuunanen, T., Rothenberger, M. a., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. http://doi.org/10.2753/MIS0742-1222240302

Phan, D. D., & Vogel, D. R. (2010). A model of customer relationship management and business intelligence systems for catalog and online retailers. *Information & Management*, *47*(2), 69–77. http://doi.org/10.1016/j.im.2009.09.001

POP, S. (2015). *2015 Email Marketing Metrics Analysis Email Marketing Metrics Analysis*.

Punera, K., & Merugu, S. (2010). The Anatomy of a Click : Modeling User Behavior on Web Information Systems Categories and Subject Descriptors. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*, 989–998.

Raj, N., Dey, L., & Gaonkar, B. (2011). Expertise Prediction for Social Network Platforms to Encourage Knowledge Sharing. *IEEE/WIC/ACM*, 380–383. http://doi.org/10.1109/WI-IAT.2011.93

Regelson, M., & Fain, D. C. (2006). Predicting Click-Through Rate Using Keyword Clusters. *Proceedings of the Second Workshop on Sponsored Search Auctions*.

Richardson, M., Dominowska, E., & Ragno, R. (2007). Predicting Clicks : Estimating the Click-Through Rate for New Ads. *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*, 521–529.

Romero, C., Ventura, S., Zafra, A., & Bra, P. De. (2009). Applying Web Usage Mining for Personalizing Hyperlinks in Web-Based Adaptive Educational Systems. *Computers & Education*, *53*(3), 828–840. http://doi.org/10.1016/j.compedu.2009.05.003

Schäfer, K., & Kummer, T.-F. (2013). Determining the performance of website-based relationship marketing. *Expert Systems with Applications*, *40*(18), 7571–7578. http://doi.org/10.1016/j.eswa.2013.07.051

Sculley, D., Malkin, R., Basu, S., & Bayardo, R. J. (2009). Predicting Bounce Rates in Sponsored Search Advertisements. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*.

SEMPO. (2012). *State of Search Marketing Report 2012 In association with SEMPO*.

Sheppard, J. S. (2013). *Factors That Influence College Choice: Decisions of Graduate Students*.

Song, Q., & Shepperd, M. (2006). Mining web browsing patterns for E-commerce. *Computers in Industry*, *57*(7), 622–630. http://doi.org/10.1016/j.compind.2005.11.006

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explor*, *1*(2), 12–23.

Tabibian, M., Amindavar, H., & Ave, H. (2001). E-COMMERCE : TECHNICAL AND

MARKET APPROACH. *International Conference on Trends in Communications, EUROCON'2001*, *1*, 149–152.

Tareef, F., & Balas, A. (2009). Marketing In Higher Education Institutes: Using An Internal Marketing Strategy And Innovation Models.

University, D. S. (2015). Dakota State University Privacy Info.

Vellingiri, J., & Pandian, S. C. (2015). A Survey on Web Usage Mining. *International Research Journal of Engineering and Technology*, *2*(1), 6–12.

Wainer, J., Dabbish, L., & Kraut, R. (2011). Should I Open this Email ?: Inbox-Level Cues , Curiosity, and Attention to Email.

Watson, T., Pay, K., Fidura, S., & Quist, D. (2013). *Email Engagement Often talked about , never defined*.

Wen, J.-R., Nie, J.-Y., & Zhang, H.-J. (2002). Query clustering using user logs. *ACM Transactions on Information Systems*, *20*(1), 59–81. http://doi.org/10.1145/503104.503108

White, R. W., Dumais, S. T., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. *ACM*, 132. http://doi.org/10.1145/1498759.1498819

Yang, Q., Va, C., & Zhang, H. H. (2001). Mining Web Logs for Prediction Models in WWW Caching and Prefetching. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)*, 473–478.

Zarrella, D. (2014). *Science of Email Marketing*.

Zhang, X., Cole, M., Street, H., & Belkin, N. J. (2011). Predicting Users ' Domain Knowledge from Search Behaviors. *ACM*, 2004–2005.

Zhu, Y., Wang, G., Yang, J., Wang, D., Yan, J., Hu, J., & Chen, Z. (2009). Optimizing search engine revenue in sponsored search. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '09*, 588. http://doi.org/10.1145/1571941.1572042