CHAPTER 3: METHODOLOGY

3.1 Introduction

The purpose of the study was to determine whether popular internet bookmarking tags can be recreated through crowd-sourcing. Amazon Mechanical Turk, the work marketplace for tasks that require human intelligence, was used as a mean to conduct the study. The study was comprises of multiple iterative experiments that were designed to achieve the highest possible quality in popular tag reproduction. Delicious - an online service for tagging, saving, and sharing bookmarks from a centralized location, most tagged websites and their tags were used as the golden set of tags to be ultimately reproduced in this study. Key research questions for the study were examined along with a number of factors regarding tag creation including the effectiveness of crowd-sourcing in reproducing popular tags, categorizing which tags can be recreated most effectively, and the relationship of worker characteristics and demographics on the effectiveness of producing popular tags.

Based on these criteria, a quantitative quasi-experimental research design was deemed to be appropriate. This chapter presents a discussion of the following specifications: (a) the research design, (b) sample size, (c) research questions/hypotheses, (d) variables, and finally (e) the data analysis that would be conducted in order to comprehensively address the research objectives. A summary will conclude the chapter.

3.2 Research Design

This proposed quantitative approach with a quasi-experimental correlational research design primarily examined whether or not popular bookmarking tags can be recreated through crowd sourcing. The main purpose of the research design is to provide a method that allows for effective and efficient reproduction of popular tags using crowd-sourcing. To this end a number of experiments were conducted. Each experiment provided useful data that suggested modifications to improve the experimental design of the study, which helped improve tag recreation activity.

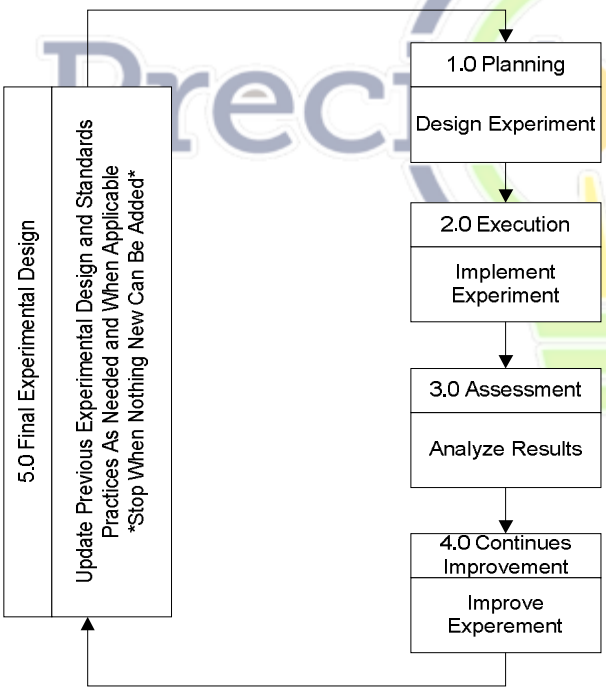

Figure 4. Iterative experimental design approach used in this study

The effectiveness of crowd sourcing in reproducing popular tags was examined using a) quantitative data derived from online surveys and b) popular tags for most tagged websites on delicious. Participants were gathered by posting tagging tasks on Mechanical Turk. Each participant was required to go through a qualification survey before he/she was trusted to take part in the research study. This quality assurance step was necessary to protect against automated scripts and workers that were trying to game the systems. There were three main objectives to the quality assurance step: a) verifying that the participants understood the task and what was requested of them; b) identifying incomplete responses or non-sense responses, c) identifying cheaters and preventing them from participating in the study. Five websites were considered for tagging tasks and this include You Tube, Flickr, Pandora, Facebook, and Digg. Those sites were chosen because they are the most all time tagged sites on delicious according to popacular.com. Popacular.com is an online service that tracks most tagged web pages on delicious at the following intervals: hourly, 8 hours, day, week, month, and all time.

The top 10 most popular tags for each one of these sites were used in this study along with data collected from the study participant survey responses. The top 10 popular tags were used as a golden set to measure the participant's ability to reproduce the same tags and exploring tag creation effectiveness with a number of user related factors. The

analysis of these variables with respect to the objectives of the study was completed by employing analysis of variance (ANOVA) and multiple linear regression.

3.3 Appropriateness of Design

The use of a quasi-experimental research design allowed the determination of whether there were statistically significant differences between groups (Cozby, 2001) in which for this study are the different tag and websites. The quasi-experimental design was appropriate to assess these differences because it allowed the researcher to compare the levels or categories of the independent variables with regard to the dependent variable in order to determine whether there was a difference between the groups (Broota, 1989).

More so, this quasi experimental correlational quantitative study specifically investigated the relationship of tagging experience (both usage and creation), search engine experience, interest in the website, and average daily time spent on the Internet of the participants. With such objective then a correlational design was appropriate. In the context of social and educational research, correlational research is used to determine the degree to which one factor may be related to one or more factors under study (Leedy & Ormrod, 2005).

The research design is quantitative for the reason that a comparison was made between an independent variable and dependent variable (Creswell, 2009). This means

that the researcher was able to quantitatively assigned numerical values to the independent and dependent variables so that a comparison was possible.

The quantitative research approach was more appropriate for this research study than a qualitative design because with a qualitative design the researcher would not be able to assess a direct relationship between two variables as result of the open-ended questions (Creswell, 2009). Qualitative design is more appropriate for observational or exploratory research that requires open ended questions and possibly ethnographic procedures. This study however follows a traditional deductive approach by building on existing theories and opertionalizing variables derived from previous empirical studies. In this study quantitative research methods are most appropriate since the researcher was able to measure the variables needed for this study and define specific research questions derived from existing research. Therefore, the quasi-experimental design was used since this would allow the researcher to determine whether there was a difference in the different tags and websites based on the dependent variables.

In order to determine whether there was a difference between the tag creation effectiveness and the various sites in terms of the tagging experience (both creation and usage), search engine experience, and average daily time spent on the Internet, an analysis of variance (ANOVA) was implemented. The ANOVA was appropriate because

the purpose was to determine whether there was a statistically significant difference between two independent populations (treatment vs. control) (Moore & McCabe, 2006). In addition, a multiple linear regression analysis was used to determine the relationship between the independent and dependent variables. The dependent variable would be tag creation effectiveness. The independent variables were interest in the website, familiarity with website, previous tag usage experience, previous tag creation experience, experience with search engines, time spend on the internet, and tag types. A multiple linear regression is appropriate because there would be multiple independent variables and only one dependent variable (Moore & McCabe, 2006).

3.4 Research Questions

A number of empirical studies concluded that social bookmarking tags can provide additional data to search engines that were not provided by other sources and consequently improve web search. The same studies however concluded that there was a lack of availability and distribution of the tags that can improve search. This study was focused on finding a way to create social bookmarking tags efficiently and effectively using crowd-sourcing.

The research questions and hypothesis that guided this study were:

RQ_1: Are there statistically significant differences in tag creation effectiveness for popular tags among the sites included in this study?

$H1_0$: There are no statistically significant differences in tag creation effectiveness for popular tags among the sites included in this study.

RQ_2: Are there statistically significant differences in tag creation effectiveness across tag types?

$H2_0$: There is no statistically significant difference in tag creation effectiveness across system types.

RQ_3: What is the relationship among tag creation effectiveness and Interest in the Website Topic, Experience with Website or Similar Website, Tagging Creation Experience, Tag Usage Experience, Experience with Search Engine, and Time spent on the Internet?

$H3_0$: None of the independent variables of Interest in the Website Topic, Experience with Website or Similar Website, Tagging Experience, Experience with Search Engine, and Time on the Internet have a statistically significant effect on tag creation effectiveness.

3.5 Population

The participants for this study were selected by posting tagging tasks on Mechanical Turk. All participants were subjected to an initial qualification survey before being allowed to participate in this study. Information related to tagging tasks was collected from the participants and were subjected for analysis.

3.6 Sampling

When calculating the sample size for the study, there are several factors that have to be taken into consideration. These factors include the power, the effect size, and the level of significance of the study. The statistical power is based on the probability of rejecting a false null hypothesis. As a general rule of thumb, the minimum power of a study that would be necessary to reject a false null hypothesis would be equal to 80% (Keuhl, 2000).

The next important factor is the effect size. The effect size is a measurement of the strength of the relationship between the independent and dependent variables in the analysis (Cohen, 1988). In most instances, the effect size of the study can be divided into three different categories: small, medium, and large.

Finally, the last two important considerations for the correct calculation of the sample size are the level of significance and the statistical procedure. The level of significance is usually set at an alpha equal to a 5% level of significance, which is

typically the standard for statistical significance. The statistical procedure must also be taken into account. Simple *t*-tests require a smaller sample than multiple regressions and, as a result, the most complicated method determines the sample size. In this case, multiple linear regression was used. Based on this information, the minimum sample size required for this study was 74 (specified as a medium effect size, a power of 95% and a level of significance equal to 5%). However in this study the overall number of participants gathered and user for the analysis was 107.

3.7 Instrumentation and Data Collection

The information that was used for this study comes from two sources:

1. Popacular.com was used to obtain the top 5 most tagged web pages on delicious. In this study the researcher used the all time data for most tagges sites. Other options include hourly, 8 hours, daily, weekly, and monthly.
2. A survey that was presented on Mechanical Turk (see Appendix A). The survey gathered key demographical information of the participants along with information pertaining to tagging tasks. The information that was gathered from this instrument included age, gender, education level, participant's interest in site, familiarity with the site, and participant's experience with search engines, time typically spent on the Internet, tag creation and usage

experience if any. The collection of data was administered through the Mechanical Turk system.

The researcher used iterative survey research design and kept updating the survey and qualifications requirements until the desired quality was achieved. There were three total iterations of this survey. Each iteration provided tags that overlap more with the golden set of popular tags gathered from popacular.com. The researcher found that the forth iteration did not provide any tag quality benefit and decided to lock in the design and instructions of the third survey. Mechanical Turk allowed workers to comment on tasks and provide feedback to requesters. The researcher found this feature to be very useful as it helped the researcher quickly identify ambiguous questions and task instructions and improve them in relatively short period of time. The initial survey design yielded low quality responses for the reason that turkers try to game the system by attempting to complete high number of human intelligent tasks (HITs) in the shortest possible time. The original task that was given to the selected participants was priced at $0.02 or 2 cents.

The initial survey did not have a qualification requirement. So in the second iteration the researcher added a qualification requirement for the available HITs. The qualification requirements were mainly geared to ensure that the workers is invested in

the task and intended to perform it well. Some of these qualification requirements included questions about the "about us" section of the sites included in the study. The questions were brief but ranged from asking the worker how many images were present on a certain web page to finding a sentence and fill in the missing words in the survey for the same sentence. In this second iteration a number of workers provided feedback regarding the working on some questions or tasks. In the third iteration, the researcher introduce a survey with improved instructions and clearly stated questions. This was the last iteration that provided the highest quality results (later iterations did not add any significant improvement). The researcher at that time finalized the survey design and launched the actual study.

The final survey contained the enhanced version of the instructions, qualification requirement and the questions related to the 5 websites. The final survey HIT was priced at $0.05 or 5 cents with 50 cents bonus if the participant completed the survey and performed the task well. A participant average time to complete the survey is 15 minutes. Shortest time was 12 minutes and longest was 19. The responses were very reliable and this had resulted in the completion of the final survey responses in 5 hours.

The raw data from Mechanical Turk was then downloaded for statistical analyses. A unique identification number was assigned to each of the participants so that no

personal information was revealed or exposed (Cozby, 2001). This identification number was used to specify each participant in the study.

3.8 Operationalization of Variables

The following variables and their specifications will be used in the analysis.

Tag Creation Effectiveness (TCE): Dependent continuous variable. TCE was calculated as the proportion of the participant-created tags that are listed on the popular tag list generated by the social network users. 10 popular tags were used for each site. Each tag was given a value that represents the usage frequency of the tag by delicious users. For example, if 100,000 users used on tag1 and 50,000 users used tag2, then tag1 is assigned a higher score then tag2 which is a reflection of frequency of use. Therefore, more popular tags, i.e. employed by more users, provide a greater variance in this variable and thus, a more robust analysis.

Tag Type (TT): Independent categorical variable. Tag type was designed to categorize the type of tags created. In this case the researcher used the tag classification schema provided by Bischoff et al. 2008, which includes: Topic, Time, Location, Type, Author/Owner, Opinion/Qualities, Usage Context, and Self [add reference].

Interest in the Website Topic (Interest): Independent ordinal variable. Interest was assessed through a 2 point Likert-scale question with 1 being most interested and 0 being least interested.

Experience with Website or Similar Website (Experience): Independent ordinal variable. Experience was assessed through a 5 point likert-scale question with 4 being most experienced and 0 being least experienced.

Previous Tag Usage Experience: Independent ordinal variable. This variable was assessed through a 5 point Likert-scale question with 4 being most experienced and 0 being least experienced.

Previous Tag Creation Experience (TCX): Independent ordinal variable. This variable was assessed through a 5 point Likert-scale question with 4 being most experienced and 0 being least experienced.

Previous Tag Usage Experience (TUX): Independent ordinal variable. This variable was assessed through a 5 point Likert-scale question with 4 being most experienced and 0 being least experienced.

Experience with Search Engine: Independent ordinal variable. This variable was assessed through a 5 point Likert-scale question with 4 being most experienced and 0 being least experienced.

Average Daily Time Spent on the Internet: Independent ordinal variable. This variable was assessed through a 4 point Likert-scale question with 3 being most time and 0 being least time.

3.9 Data Analysis

The data analysis that was used in this study comprised of descriptive statistics, analysis of variance (ANOVA), and multiple linear regression. Each of these analyses was conducted in SPSS Version 16.0®.

3.9.1 Descriptive Statistics

The descriptive statistics was comprised of frequency distributions as well as measures of central tendency. For the frequency distributions, the number and percentage of each occurrence were presented for the categorical variables in the study. The measures of central tendency included the presentation of the mean, standard deviation, and minimum and maximum values for the continuous variables in the study such as the age of the participant.

3.9.2 ANOVA

As a subsequent analysis, an ANOVA was conducted for the first and second hypothesis. The ANOVA is a statistical method that is used in order to determine whether an independent variable(s) has a significant impact on a single dependent

variable. An advantage of the ANOVA is that it allows one to be able to include more than one independent variable in the model at the same time in order to determine the effect of each variable or to control for specific variables (Tabachnick & Fidell, 2001). In other words, one is not limited to only including one variable in the analysis. This is important since this allows one to control for a number of variables that may be related to the dependent variable.

When the variables have been included in the ANOVA model, the results would indicate whether an individual or several independent variables contribute to the explanation in the variation of the dependent variable (Tabachnick & Fidell, 2001). What this means is that if a variable is found to be significant then it could be concluded that this variable significantly contributes to the explanation in the variation of the dependent variable (Keuhl, 2000). The significance of the test is based on an F-statistic that is from the F-distribution (Keuhl, 2000). Therefore, if the F-statistic exceeds this critical value then one would be able to conclude that there is a relationship between the independent and dependent variables.

3.9.3 Multiple Linear Regression

A multiple linear regression model was used specifically for the third research question. The dependent variable would be tag creation effectiveness. The independent

variables were Interest in the Website Topic, Experience with Website or Similar Website, Tag Creation Experience, Tag Usage Experience, Search Engine Experience, and Time on the Internet. A multiple linear regression is appropriate because there would be multiple independent variables and only one dependent variable. This would be the most complex of the analyses because there would have to be more assumptions made in order to make valid inferences about the target population. The one limitation to this multivariate analysis is that the regression residuals must be normally distributed. Statistically significant parameter estimates for the multiple linear regression at the 0.05 significance level would be sufficient evidence to reject the null hypothesis.

3.10 Summary

This chapter presented the type of research design that was used which is a quasi-experimental correlational design. This was chosen because it is the objective of the study to determine whether there are significant relationships between or among tag creation effectiveness and a number of independent variables. Mechanical Turk workers were surveyed and used as participants for this study. In terms of the statistical analysis, three separate statistical tests were used. Descriptive analysis, ANOVA, and multiple linear regression were deemed to be the most appropriate methodologies for testing the hypotheses of the study. This chapter also discussed the source of the data, research

questions and procedures, hypotheses and data collection. The data analysis and results will be discussed in Chapters four and five.