Characterization of Individual Health Topic Familiarity in Consumer Health Information Search

Submitted to
Graduate School of Information Science and Technology
Osaka University

April 2015

Ira PUSPITASARI

Publications

Journal Paper

1. Ira Puspitasari, Koichi Moriyama, Ken-ichi Fukui, and Masayuki Numao. “Effects of Individual Health Topic Familiarity on the Activity Pattern during Health Information Searches”, *JMIR Med Inform* 2015; 3(1): e16, doi:10.2196/medinform.3803. PMID: 25783222

International Conference and Workshop Papers

1. Ira Puspitasari, Ken-ichi Fukui, Koichi Moriyama, and Masayuki Numao. "Predicting Consumer Familiarity with Health Topics by Query formulation and Search Result Interaction", *LCNS Vol. 8862, PRICAI 2014: Trends in Artificial Intelligence*, Gold Coast, Australia, Dec. 2014.
2. Ira Puspitasari, Ken-ichi Fukui, Koichi Moriyama, and Masayuki Numao. “Health Information Search Personalization with Semantic Network User Model.” *Proceedings of Workshop on Computation: Theory and Practice WCTP2013*, pp. 168-177. World Scientific, 2013.
   doi: 10.1142/9789814612883_0012
3. Ira Puspitasari, Roberto Legaspi, and Masayuki Numao. "Characterizing the Effect of Consumer Familiarity with Health Topics on Health Information Seeking Behavior", Proc. The 27th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2013), Toyama, Japan, June 2013. Available online:
   https://kaigi.org/jsai/webprogram/2013/pdf/995.pdf
4. Ira Puspitasari and Masayuki Numao. "The Framework of Evolutionary Community of Practice." *Knowledge Co-Creation 2* (2012). Available online:
   http://www.jaist.ac.jp/fokcs/2012/publications/FOKCS2012Mar_Final_PUSPITASARI%20.pdf

Presentations

1. "Predicting Consumer Familiarity with Health Topics by Query Formulation and Search Result Interaction" *The 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI) 2014*, Gold Coast, Australia, December 1- 5, 2014.
2. "Personalizing Health Information Retrieval System by Context and Health Topic Familiarity", Poster presentation, *The 17th SANKEN International Symposium*, Osaka, Japan, January 21-22, 2014.
3. "Personalization Approach in Health Information Retrieval System", *Type II Presentation The 2013 International Conference on Active Media Technology & Brain and Health Informatics*, Maebashi, Japan, October 29-31, 2013.
4. "Health Information Search Personalization with Semantic Network User Model", *Workshop on Computation: Theory and Practice (WCTP) 2013*, Manila, Philippines, September 30 – October 1, 2013.
5. "Personalization Approach in Health Information Retrieval System", *The 4th International Workshop on Empathic Computing (IWEC) 2013*, Beijing, China, August 3 – 5, 2013.
6. "Characterizing the Effect of Consumer Familiarity with Health Topics on Health Information Seeking Behavior", *The 27th Annual Conference of the Japanese Society for Artificial Intelligence*, Toyama, Japan, June 4 – 7, 2013.

Abstract

The emergence of e-patient has encouraged non-medical professionals (consumers) to be more proactive regarding healthcare education and health decision making. However, searching for understandable health information on the Internet is challenging for most consumers that have different health topic familiarities. A consumer could be knowledgeable about *skin allergy* but uninformed about *heart attack*, whereas another consumer may have the reverse health topic familiarities. The term *diabetes mellitus* may be well understood by some consumers, but completely unfamiliar to other consumers. This variation in familiarity may cause misunderstandings because the information presented by health information search systems may not fit the consumer's understanding.

This research aims to design and develop individual health topic familiarity concept as the determinant factor in personalizing health information search systems. The first research work is to examine the effects of health topic familiarity on health information search behaviors. For this purpose, we defined three categories of health topic familiarity, i.e., unfamiliar (L1), somewhat familiar (L2), and familiar (L3). The analysis of state transitions in search activities detects unique behaviors and common search activity patterns in each familiarity group. The most common patterns in group L1 were frequent query modifications, with relatively low search efficiency, and accessing and evaluating selected results from a health website. Group L2 performed frequent query modifications, but with better search efficiency, and accessed and evaluated selected results from a health website. Finally, the members of group L3 successfully discovered relevant results from the first query submission, performed verification by accessing several health websites after they discovered relevant results, and directly accessed consumer health information websites.

The next research work is to extract the features set from the identified unique behaviors and to develop a familiarity prediction model based on these features. The extracted features set are the query formulation and search result interaction. The results show that the prediction model achieved high accuracy, within 80% - 90%, in identifying consumer's health topic familiarity. This finding suggests that health topic familiarity identification based on the query formulation and the search result interaction is feasible and effective.

Acknowledgements

الْحَمْدُ للَّهِ رَبِّ الْعَالَمِيْن

Alhamdulillāhirobbil'ālamīn.

I have completed this work under the supervision of Professor Masayuki Numao. I express my sincere gratitude to him; his patient guidance and support over the past three years that teaches me many important aspects of research works. It has been a great pleasure to be one of his students.

I would like to express my gratitude to Directorate General of Higher Education, Republic of Indonesia for the generous support and scholarship during my study in Osaka University.

It is a fact that the completion of my work is possible due to the support of many individuals. It is too numerous to name all but I at least make an attempt.

There is Febdian Rusydi, my husband, for his genuine devotion and excellent support. Then my younger brothers, Rendra Adi Kusuma and Rendra Faris Ramadhan, for always standing by my side whatever my situation is.

There are Professor Koichi Moriyama and Professor Ken-ichi Fukui for going to great lengths to provide guidance and insights to my research. Then Professor Satoshi Kurihara and Dr. Roberto Legaspi for helping me in developing and improving my research. I also thank Professor Merlin Suarez of De La Salle University, The Philippines for her valuable support and helping my data collection.

There are Professor Atsushi Yagi (Division of Nonlinear Systems, Modeling and Optimization), Professor Hiroshi Morita (Division of Systems Engineering), and Professor Yasumasa Fujisaki (Division of Operations Research) for reviewing my work in details. I am truly grateful for all the questions, suggestions, and comments that improve my work.

Then the students of Numao Laboratory for their help, support, and friendship. My great thanks goes to everyone in this lab, former and present. I thank the laboratory secretaries, Ms. Misuzu Yuuki, Ms. Megumi Tanabe, Ms. Azusa Hirabayashi and Ms. Mika Kusakabe for helping me in all administration matters.

And, last but not least, my home institution, Universitas Airlangga (Surabaya, Indonesia) for allowing me to take a special leave for my education. I also thank my colleagues in Information Systems department for their great supports.

I dedicate this work to my beloved parents,
Mochamad Indra and Reny Angraeny

This page is intentionally left blank.

Table of Contents

Publications ..... **i**

Abstract ..... **iii**

Acknowledgements ..... **v**

Table of Contents ..... **ix**

List of Figures ..... **xiii**

List of Tables ..... **xv**

Chapter 1 Introduction **1**

1.1 Background of the Research 1

1.2 Research Objectives 3

1.3 Significance of the Research 4

1.4 Research Methodology 5

1.5 Structure of the Thesis 6

Chapter 2 Literature Review of Related Work **9**

2.1 Interpretive Layer in Health Information Search System 9

2.2 User's Interaction Behavior on Web Information Search 10

2.3 Familiarity Concept in Web Information Search 13

Chapter 3 Health Topic Familiarity **15**

3.1 Category of Health Topic Familiarity 15

3.2 The Importance of Consumer's Health Topic Familiarity in Health Information Search 17

3.3 Factors Characterizing Health Topic Familiarity in Health Information Search 17

3.3.1 Search Activity....18
3.3.2 Query Formulation and Search Result Interaction....21

Chapter 4 Experimental Design Supporting Individual Health Topic Familiarity in Health Information Search....23

4.1 Instrument....23
4.1.1 Health Terminology Familiarity Questionnaire....23
4.1.2 Health Search Task....24
4.2 Data Collection Procedure and Data Analysis....26
4.3 Demographic Profile of the Participants....27
4.4 Health Topic Familiarity Questionnaire Result....28

Chapter 5 The Effects of Health Topic Familiarity on Health Information Search Behavior....31

5.1 Method of Examining the Effects of Health Topic Familiarity on Search Behavior....31
5.1.1 Modeling the Search Session....32
5.1.2 Calculating the Transition Frequency between Search Activity Types....33
5.1.3 Identifying Search Activity Patterns....34
5.2 Result....35
5.2.1 Frequency of Search Activities....35
5.2.2 Transition between Search Activity Types....37
5.2.3 Testing the Differences in the Search Activities between Familiarity Groups....39
5.2.4 Most Frequently Pattern of Search Activities Sequence Applied in Each Familiarity Group....40
5.3 Analysis and Discussion....44

Chapter 6 Prediction Model of Health Topic Familiarity based on Health Information Search Behavior....49

6.1 Features....49
6.2 Features Selection....53
6.3 Model Development....54
6.4 Result and Analysis....55

Chapter 7 Summary 59

7.1 Conclusion 59

7.2 Implications for Health Information Search System Design 60

7.3 Limitations and Future Studies 60

References 63

Appendix A Health Terminology Familiarity Questionnaire 71

Appendix B Transition between Search Activity Types 77

Appendix C Most Frequent 5-gram Sequence Patterns in Each Familiarity Group 81

This page is intentionally left blank.

List of Figures

Figure 1.1 Research Methodology 5
Figure 1.2 Structure of the Thesis 8
Figure 3.1 The difference between Category L1, L2, and L3 16
Figure 3.2 Factors characterizing health topic familiarity in health information search behavior 18
Figure 4.1 Examples of the questions included in the health terminology familiarity questionnaire 24
Figure 5.1 The method of examining participant's search behavior in health information search 31
Figure 5.2 Percentage of the search activity types in all familiarity groups .. 35
Figure 5.3 Perplexity values for L1, L2, L3, and all the test data using different n-gram models 40
Figure 5.4 Comparison of frequent activity patterns in Category 1 41
Figure 5.5 Comparison of frequent activity patterns in Category 2 42
Figure 5.6 Comparison of frequent activity patterns in Category 3 42
Figure 5.7 Comparison of frequent activity patterns in Category 4 42
Figure 6.1 The extraction scheme from Search Activity Types to Features Sets 50
Figure 6.2 Model development 55

This page is intentionally left blank.

List of Tables

Table 3.1 Stages and search activity types 19
Table 4.1 Health search tasks 25
Table 4.2 Demographic profile of the participants 28
Table 4.3 The familiarity questionnaire result 29
Table 5.1 Frequency and proportion of search activity type 36
Table 5.2 Top 10 frequent first order transitions for each familiarity group 38
Table 5.3 Results obtained after testing the differences between the familiarity groups ($P < .001$) 39
Table 5.4 Comparison of frequent activity patterns 41
Table 5.5 Summary of the findings from the experiment in Chapter 5 47
Table 6.1 The query formulation features 50
Table 6.2 Query type definition 51
Table 6.3 The search result interaction features 52
Table 6.4 Information gain of each feature 53
Table 6.5 Accuracy of the classifiers 56

This page is intentionally left blank.

Chapter 1

Introduction

Chapter 1 describes the research background in this thesis, the objectives and the significances of the research, the research methodology, and overview of the thesis structure.

1.1 Background of the Research

The e-patient movement has emerged the awareness of health information literacy among the people of non-medical professionals (consumers). The consumers are the patient, the patient's family and caregiver, and the people who occasionally search for general medical health and wellness information. More consumers are progressively using the Internet to support health information needs [1-5]. A number of support systems have been developed to provide access to consumer-friendly health information. However, searching for understandable health information on the Internet is difficult for most consumers because they are not familiar with the standard terminology employed in healthcare publications [6-9]. Difficulties arise when formulating queries and when trying to understand the health information presented.

Researchers and healthcare providers are working on consumer-based initiatives to resolve the communication gap problem. Soergel et al. [9] proposed an "interpretive layer" design to assist consumers when formulating effective queries, finding and interpreting relevant health information, and applying the information in an appropriate manner. This interpretive layer design concept has been implemented in several consumer health systems, such as Interactive Online Health Information System [10], Query Assistant in Health Information Search System [11], MedicoPort [12], and MedSearch [13]. To further reduce the communication gap between consumers and healthcare professionals/health materials, several researchers have studied the familiarity and recognition rate of

health terminologies among consumers [6-7, 14-17], and developed automated tools for assessing the readability of health texts [18-19].

Most studies of health information search by consumers have focused on improving the health search experience of consumers by providing intelligent assistance and utilizing more consumer-friendly terminology. However, there is a lack of research on *individual* health topic familiarity and how this familiarity influences health information search behaviors in *specific* consumers. The familiarity with health topic affects the search process (e.g., the chosen search strategy/tactics, the performed search activity pattern, the submitted query, and the visited retrieved search results) and the search outcome (i.e., the quality of health information found by the searcher). These research topics are important because every consumer has different health topic familiarities as in the following cases:

1. A consumer is familiar with several health topics, e.g., *hypertension*, *cholesterol problems*, and *diabetes*, but he/she is unfamiliar with other topics.
2. A consumer is well informed about "*skin allergy"* but uninformed about "*cardiovascular disease*," whereas another consumer may have the opposite health topic familiarities.
3. The term "*gastro-esophageal reflux disease"* may be well understood by some consumers, but completely unfamiliar to other consumers.

This diversity may lead to misunderstandings because the information presented during health information searches may not suit the consumer's familiarity. Misunderstandings in health information may lead to unwise health decisions [20] that affect a person's life.

Information in health domain varies from general article to a complicated medical report. Health information search system should be able to present health information that matches consumer's understanding as closely as possible. Introductory information about *heart attack* is relevant for the consumer who had never heard the terminology before, while *modern management of acute myocardial infarction* article is more suitable for the familiar consumer. Thus, a

personalization approach in based on the consumer's familiarity in health information search system is required to avoid misunderstanding and to improve the overall search process.

1.2 Research Objectives

This thesis aims to propose *individual* health topic familiarity as a determinant factor in personalizing health information search system. Consumers with different familiarities need different type of information. Consumers may use health information presented by the search systems to make health related decisions; therefore the retrieved health information should be matched as closely as possible to the consumer's level of understanding.

To accomplish the main objective, this thesis addresses the following research questions:

1. How the individual health topic familiarity affects health information search behavior?

   Health information search requires high cognitive load. It is important to examine how the familiarity with health topics influences search behaviors. Characterizing the common search behaviors exhibited by consumers with different levels of familiarity facilitates the identification of suitable system's support to improve the overall search process.
2. How to develop a prediction model of health topic familiarity based on health information search behaviors?

   Identifying the consumer's familiarity with the health topic being searched is necessary to create a personalized model for each consumer's and to deliver the result as accurately as possible. Analyzing the search behavior is one of the most preferable methods to create an optimized personalization without additional efforts from the user.

1.3 Significance of the Research

Despite the increasing number of health information search systems and the greater amount of health information on the Internet, consumers are still having difficulty acquiring and filtering proper health information. Researchers and healthcare professionals have developed consumer-friendly systems to overcome this problem. However, there is a lack of support in accommodating individual health topic familiarity in health information search systems.

This thesis focuses on health topic familiarity and contributes to:

1. The observation of individual health topic familiarity in health information search.

   Every person has a unique health topic familiarity map (list of health topics that the person is familiar with). A consumer can be well informed about certain health topics but unfamiliar to other topics. Results and findings of this research support this observation.

2. The identification of unique search pattern between different familiarity groups (unfamiliar, somewhat familiar, and familiar groups).

   Health information search system can use this knowledge to automatically identify the consumer's familiarity with health topic by analyzing the consumer's search behaviors. Then, the system creates a personalized model for each consumer, delivers relevant results, and provides suitable supports based on consumer's familiarity.

3. The development of familiarity prediction model based on the consumer's search behavior.

   As the first step toward the improvement of more consumer-friendly health information search system, the system must be able to identify the consumers' familiarity by their search behavior. We developed a familiarity prediction model based on consumer's search behavior, i.e., query formulation and search result interaction. The proposed prediction model performs reasonably well with 80 – 90% accuracy.

1.4 Research Methodology

The research work started with review of the related literature on consumer health informatics, health information search systems, user's interaction behavior, study of user's familiarity, and background knowledge in web information search. The purpose of the literature review was to identify the remaining major problem in health information search, limitations of the current solutions, and what approach can be used to improve these limitations and to solve the research problems.

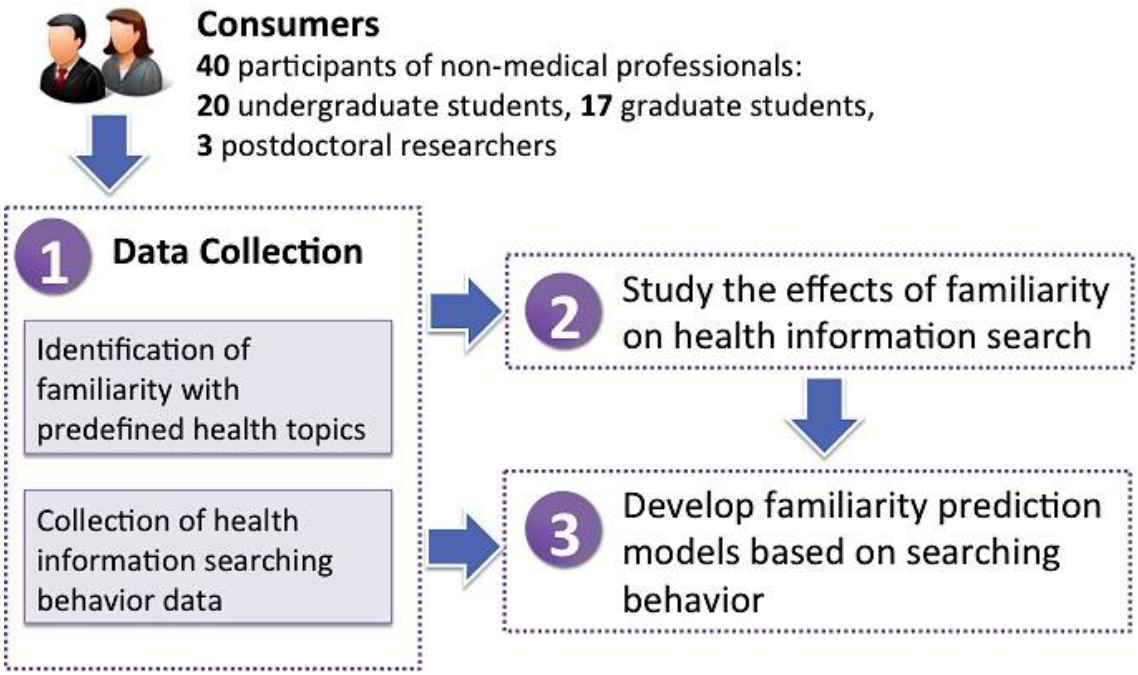


Figure 1.1 Research Methodology

The methodology comprised of three main research activities as shown in Figure 1.1: data collection, the study of health topic familiarity effects on health information search, and the development of health familiarity prediction model based on consumer's searching behavior. Data was collected from 40 participants of non-medical professionals (students and postdoctoral researchers). A complete data collection consisted of demographic profile survey, health terminology familiarity questionnaire, and health information search sessions. The next activity, examining the effects of health topic familiarity, was aimed to identify the

characterization of health information search based on topic familiarity. The finding from this activity was used to develop the familiarity prediction model based on the characterization of health information search behavior.

1.5 Structure of the Thesis

Chapter 1 describes the background and the identification of the remaining major problem in consumer health information search. The chapter includes the explanation of the research objective, significance of the research and the detail methodology.

Chapter 2 provides state of the art of the related work that is applied through this thesis. The chapter begins with the recent survey in interpretive layer in health information search and user's interaction behavior in web information search. The chapter ends with the review on familiarity concept, which lays the foundation of the individual health topic familiarity concept proposed in this thesis.

Chapter 3 describes the health topic familiarity concept proposed in this thesis as the solution of the research problem. This chapter includes the definition of health topic familiarity, the classification of health topic familiarity, and the factors characterizing health topic familiarity in health information search.

Chapter 4 describes the experimental design of the research. This chapter begins with the explanation of the data collection instrument and the procedure of data collection. The demographic profiles, and the health topic familiarity of the participants are also presented in this chapter.

Chapter 5 is to answer the first research question. This chapter discusses the importance of health topic familiarity in health information search process, the detail explanation of the method employed, the result, and the analysis of the result.

Chapter 6 is to answer the second research question. This chapter discusses the development of prediction model of individual health topic familiarity based on health information search behavior. It includes the detail explanation of the model development process, features selection, the prediction model performance, and the analysis of the result.

Chapter 7 concludes this thesis. This chapter also describes the limitation of the current study in this thesis and suggests future improvement.

The schematic diagram of the thesis structure is presented in Figure 1.2.

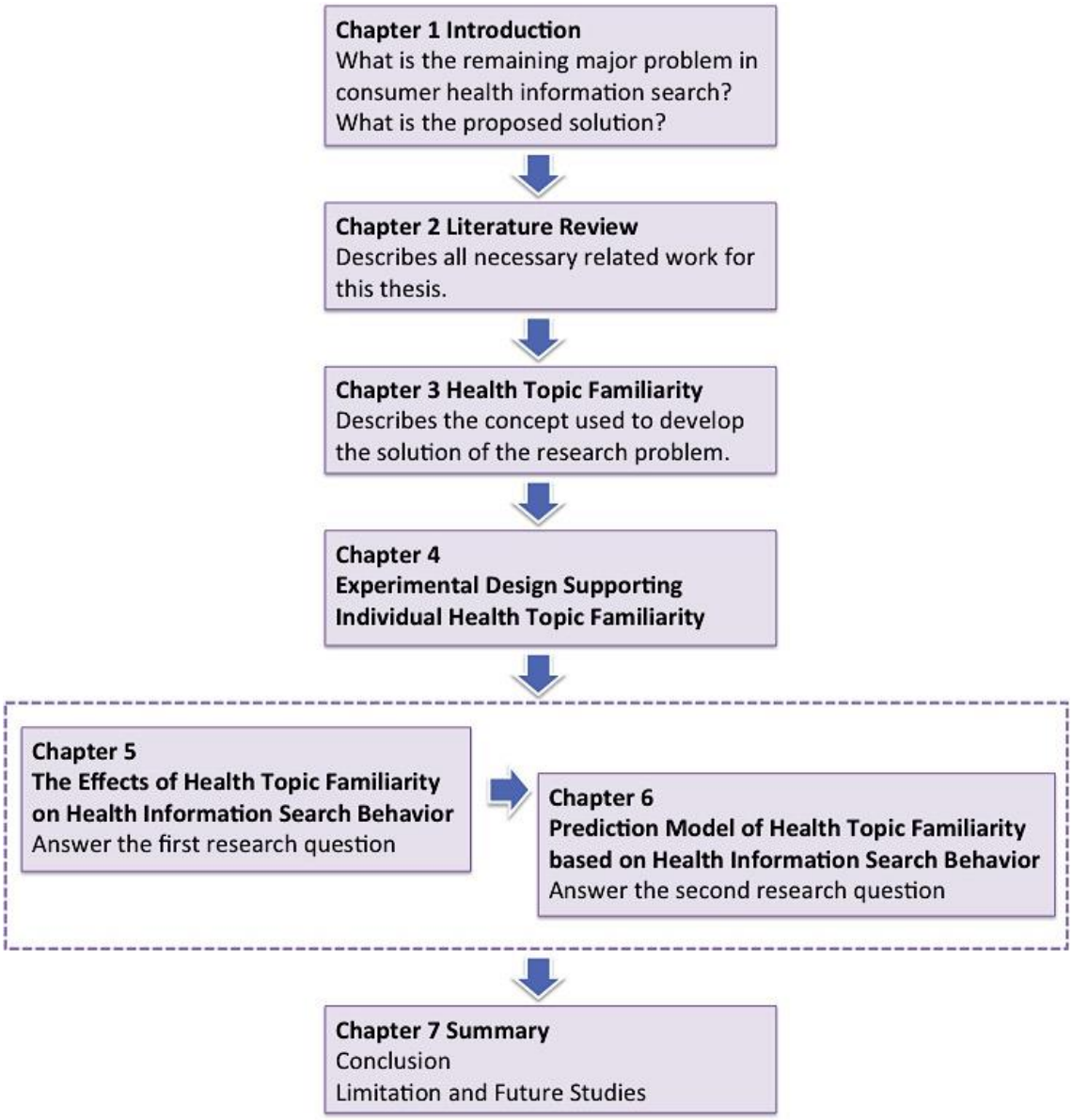


Figure 1.2 Structure of the Thesis

Chapter 2

Literature Review of Related Work

This chapter provides an overview of the related work in health information search system, user's interaction behavior in web information search, user model based on background knowledge and search topic familiarity, and familiarity concept in web information search.

2.1 Interpretive Layer in Health Information Search System

The interpretive layer framework in consumer health information search system was proposed to bridge the gap between consumer understanding and biomedical knowledge at all levels [9]. According to Soergel et al., this layer aims to help consumers in understanding their medical problems, formulating effective queries, navigating the systems, understanding the documents found, and applying the information appropriately. Researchers and health care professionals have developed consumer-friendly health information systems based on the interpretive layer framework.

Some of the health information search systems that specialize in assisting the consumers to better understand their health problems are MedSearch [13] and Intelligent Medical Search Engine (iMed) [21]. MedSearch is a medical specialized search engine that accepts long queries in plain English. The search engine extracts the representative keywords from the submitted query. Based on the extracted keywords, MedSearch returns diversified search result and suggests related medical phrases with proper ranking and annotation. These features were built based on the behavior of general consumers, which prefer to formulate readable long query and to receive all kinds of medical knowledge related to their situation. The next system, iMed, uses predefined questionnaire to capture consumer's health information need. Based on the questionnaire response, iMed automatically

forms the medical query; structures the entire search results into multilevel hierarchy; and suggests related medical phrases.

The systems that implemented the second function in interpretive layer framework are Health Information Query Assistant (HIQuA) and MedicoPort. HiQUA provides alternative query terms related to the consumer's initial query [11]. The suggested terms are selected based on their semantic distance from the original query and the co-occurrences in medical literature and log data. The next system, MedicoPort, uses Unified Medical Language System (UMLS) resources to increase the effectiveness of medical search for non-medical professional searchers. Its query formulator and concept generator of MedicoPort uses UMLS Metathesaurus and UMLS Semantic Network to expand user query, reformulate query terms, rank the search result, and filter irrelevant documents [12].

To help consumers obtain understandable and relevant health information based on their needs, some health information search systems have added the personalization feature. The early personalization approach is integrating electronic patient records with health-related content on the Internet. The project of Structured Evaluated Personalized Patient Support uses electronic patient data to construct user profiles and to retrieve health information based on the profiles [22]. The next approach in the personalization of health information search system is using the user-centered design concept. Le Rouge et al. applied this methodology to design the Consumer Health Technologies device for aging population who suffered from diabetes [23].

2.2 User's Interaction Behavior on Web Information Search

Information seeking on the Internet is an interactive and iterative process [24], and a learning process [25]. In Saracevic's stratified model, the interaction between users and systems in an information retrieval system occurs in several connected levels [26]. There are several levels on both sides, i.e., cognitive, affective, and situational levels on the user side; and engineering, processing, and content levels on the system side. The user and the system meet via an interface on

the surface level. User performs search strategies, submits a query, and selects potential relevant documents returned by the system. All user actions are the reflections of the cognitive, affective, and situational connected levels. However, the problem in most information retrieval system is the system fails to understand the deeper levels of the user. Most of the interaction only occurs on the surface level; therefore the system outcome is not suitable with the user's needs. Researchers have proposed solutions to this problem by analyzing the user's interaction behavior during a search session, such as the query formulation and reformulation [25, 27-29], the selection of potential relevant results [30, 31], and the search strategies and tactics [32-35].

Query formulation and reformulation has been considered as one of the most essential interactions between users and information retrieval systems [25, 27, 35]. Rieh and Xie examined the sequence of multiple queries because the query reformulation expressed the deeper level of the interaction on the user side. They proposed a model of web query reformulation patterns, i.e. specified reformulation (specify the meaning of subsequent query by adding more terms or replacing terms with more specific meaning terms), generalized reformulation (generalize the subsequent query by deleting terms or replacing terms with more general meaning terms), parallel reformulation (modifies the queries from one aspect to another, which share common characteristics), and building block reformulation (identify and combine multiple concepts from the previous queries and use them in subsequent queries), dynamic reformulation (employ inconsistent pattern, move around from one type to another type), multi-tasking reformulation (search for two or more topic in the same search session), recurrent reformulation (resubmit the exact same query from the previous queries), and format reformulation. In another study, Boldi et al. classified query reformulations using two dimensions taxonomies: the generalization-specialization axis and dissimilarity axis [36]. The first axis depicts the reformulation between more general and more specific query, while the second axis portrays the change in syntactic and semantic between two queries from Same Query, Error Correction, Equivalent Rephrasing, Parallel Move, to Mission Change pattern. To improve the

user interaction in information retrieval system, the system needs to support various kinds of query reformulation patterns. Some other studies in query formulation and reformulation proposed other important factors in characterizing the user behavior, such as query length and query vocabulary [30], quantitative query attributes [37], and cognitive styles [38].

The next important interaction in information search is identifying and selecting the potential relevant results. White et al. (2009) investigated the source of web sites visited by domain expert and non-expert users. Expert users were likely to visit specific technical web sites, while non-expert users were interested in consumer-oriented or advisory web sites. For example, the computer science experts visited specific programming language web sites, while the non-experts were more concerned with general computer topics. Researchers also examined page dwell time and reading level of the visited web sites to characterize the search behavior [39-41].

Researchers have also studied the search strategies employed by the users during information seeking on the web environment. Search strategy consists of search tactics and moves [32, 35], including tactics in query formulation and reformulation [42], and interaction with the search results [43]. One of the techniques applied to study the search strategy is analyzing the sequential transition from one search tactic to another search tactics. Wildemuth (2004) investigated the effect of domain knowledge in searching behavior by applying maximal repeating patterns (MRPs) analysis to the sequence of search tactics moves [27]. This study revealed that participants with lower domain knowledge performed less efficient search moves with more errors occurred in the reformulation of the search tactics. Xie and Joo (2010) examined user's transitions in search tactics during a Web-based search process [35]. They applied fifth order Markov chain to discover the most common patterns of search tactic transitions at different phases within a search session. In another study, Lin and Wilbur (2009) modeled the sequence of user actions in PubMed search engine using n-gram language model [44].

2.3 Familiarity Concept in Web Information Search

One of the most useful features to model the user as an individual is the user's domain knowledge or background knowledge. Domain knowledge refers to the user's knowledge of the subject area that is the focus or topic of search [27, 30]. Other studies used the terminology familiarity to express the user's domain knowledge and how it affects the search process [27, 30, 45, 46]. Users who have greater familiarity with the search topic use more varied and specific vocabulary [30, 47], perform specific and advanced search strategies [27, 30, 32, 37, 45], and have better search efficacy [27, 30].

Previous studies in the information search area have demonstrated the impact of topic familiarity on search behaviors [27, 29, 30, 45, 46]. Searchers who have greater familiarity with the search topic use more varied and specific vocabulary [30], perform specific search strategies [30, 45], and have better search efficacy [27]. One approach for examining search behaviors is to analyze the search activities performed by seekers [27, 45]. Several studies have addressed the activities involved in search tactics [32] and search strategies [33, 34]. To obtain a more comprehensive understanding, researchers have also studied the transitions among states during search activities [35, 42, 48], and analyzed the sequence of search activity transitions using state transition network [27] and Markov chains [35, 44, 49].

In health information search domain, the study of familiarity is employed to reduce the communication gap between consumers and healthcare professionals/health materials. Zeng et al. developed the Consumer Health Vocabularies (CHV) initiative project, which links the vocabulary of consumers to the terminology used by healthcare professionals and in healthcare materials [6]. By building on the CHV project, several studies have proposed predictive models for measuring the average familiarity of various consumer health vocabularies based on the term occurrence in text corpora [14], demographics factors [15], and contextual features [16,17]. In attempts to provide more consumer-friendly health materials, other researchers have developed automated tools for assessing the readability of

health texts by substituting difficult terms with easier synonyms and simplifying long sentences [18] or by comparing the terms appeared in a document and terms known by the user [19]. Another study to improve the availability of consumer-friendly information is the consumer health educational project by European Patients' Academy on Therapeutic Innovation (EUPATI) [50].

Chapter 3

Health Topic Familiarity

The World Health Organization (WHO) defined health as "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity [51]." Health is an individual state that is influenced by key determinants, such as economic and social environment, education and literacy, personal health practices, and culture [51]. Some examples are:

1. Higher education level is linked with good health awareness and healthier lifestyle.
2. People who live in an outbreak area are familiar with the epidemic disease.
3. Patients and their caregivers educate themselves with the disease information (prognosis, treatment, medicine).
4. Genetics and family condition affect healthiness and the likelihood of developing certain disease.

Health topic familiarity in this study refers to the degree of acquaintance with a specific health topic. The composition of familiar health topics of each person varies depending upon individual key health determinants. A person could be knowledgeable about *typhoid fever* and *tropical disease*, but unfamiliar with *food allergy*, whereas another person has the opposite familiarities. The terminology *arthritis* may be well understood by some consumers, but completely unfamiliar to other consumers.

3.1 Category of Health Topic Familiarity

This thesis modifies and extends the familiarity types described in [15]. There are three groups of health topic familiarity defined in this thesis. The categorization is created based on the consumer's level of understanding with the key phrases defining a health terminology as follows:

1. Category L1 refers to an unfamiliar consumer who had never heard of a health terminology before or only recognized it at the surface level.
2. Category L2 refers to a consumer who has some familiarity to associate the consumer-friendly health terminology with the basic phrase defining the terminology.
3. Category L3 refers to a consumer who has a good familiarity to associate the consumer-friendly terminology and its corresponding advanced terminology with the basic phrase defining the terminology.

Figure 3.1 illustrates the difference between the categories using the terminology *heart attack*. A consumer belongs to category L1 if he/she had never heard about heart attack before or recognizes it as a condition related to *heart and blood vessel disease* (surface level). A consumer in category L2 knows the key phrases defining the terminology heart attack, i.e., the artery that carries blood the heart is blocked. By understanding the definition, the consumer in category L2 is able to distinguish heart attack with other cardiovascular disease topics, such as *heartbeat rhythm problem*, and *sudden cardiac arrest*. A consumer in category L3 acknowledges the key phrases defining heart attack and the corresponding advanced terminology.

Heart attack

Category L1

Category L2

Category L3

Surface level	Heart and blood vessel disease	Heart and blood vessel disease	Heart and blood vessel disease
Definition of consumer-friendly health terminology		the artery that carries blood to the heart is blocked	the artery that carries blood to the heart is blocked
Corresponding advanced health terminology			*Myocardial infarction* a blockage in the artery that carries blood to the heart

Figure 3.1 The difference between Category L1, L2, and L3

3.2 The Importance of Consumer's Health Topic Familiarity in Health Information Search

Health information search has specific characteristics compares to general web search. First, health information search process requires more stringent regulation because it directly affects a person's life. Misunderstanding health information may lead to *cyberchondria* phenomenon and unwise health decisions. Cyberchondria happens when minor symptoms are interpreted to serious illness [52]. Therefore, the health information presented should be matched as closely as possible with the consumer's familiarity.

The next specific characteristic is the difficulty of the terminology. In general domain, long and complicated word/phrase is, on average, more difficult than short word/phrase [53]. However in health domain, the difficulty of the terminology is not parallel with the length of the word/phrase. For example, the word *biopsy* is perceived to be more difficult than the phrase *diabetes mellitus*, and consumers are more unaccustomed to the word *dysphagia* than *hypertension*. In addition, consumer's familiarity with difficult terminology in health domain is not necessarily determined by the education level as other general domains might be. An experience with an illness may override the insufficient education level [53].

3.3 Factors Characterizing Health Topic Familiarity in Health Information Search

Consumer search behavior in a health information search session reflects his/her understanding toward the health topic being searched. It is expected that unfamiliar consumers, who had never heard the search topic before, would likely face difficulty in the search process and would take longer route to find the qualified health information. They need to build their understanding with the search topic before they can locate the relevant information. Unfamiliar searchers are also prone to unreliable health information. On the other hand, familiar searchers take advantages of their knowledge by using more precise keywords and accessing trustable sources to find the qualified health information. Figure 3.2

shows the three factors deduced from the search behavior that characterize the consumer's familiarity. The following subsections describe each factor in detail.

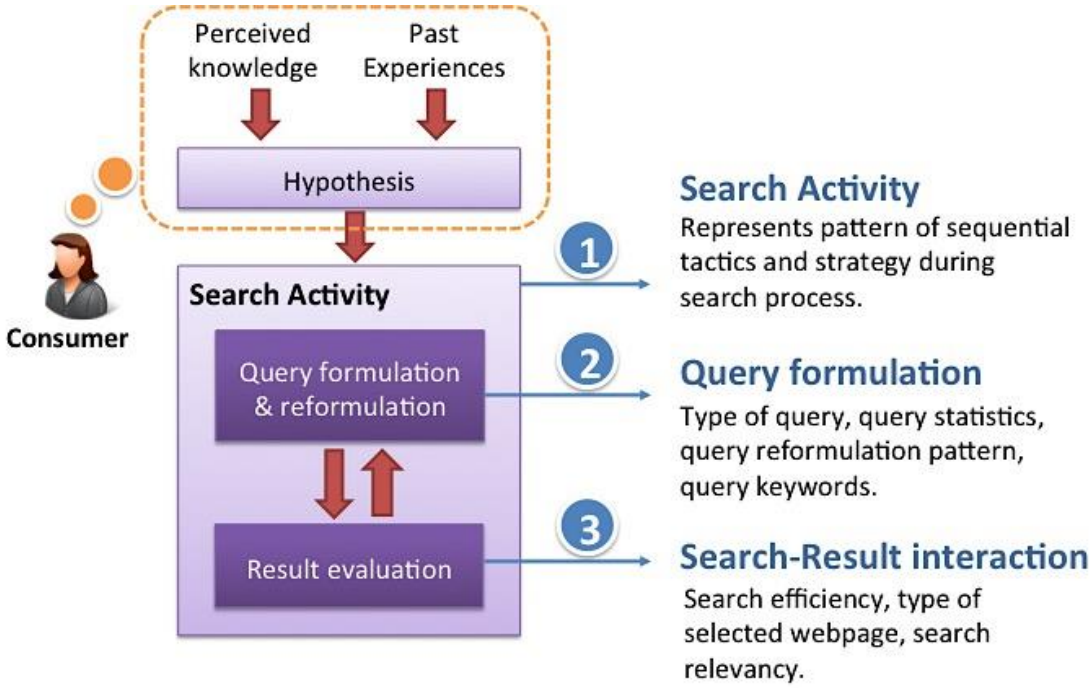


Figure 3.2 Factors characterizing health topic familiarity in health information search behavior

3.3.1 Search Activity

A search activity comprised an action, which included an operational move and a conceptual strategy that the consumers employed to achieve their goal during the health information search process. There are five stages and 18 types of search activity defined in this thesis to model the health information search session, as presented in Table 3.1. Five search activity types were modified from the study reported by Xie and Joo [35], i.e., "Examining the retrieval result (E:ExamSR)," "Evaluating the selected item (webpage) (E:EvalI)," "Exploring link forward (A:XplorF)," "Accessing link backward (A:AccB)," and "Using the information (Use)."

Table 3.1 Stages and search activity types

Stage	Search Activity Code	Description
Querying	Q:AccSE	Access a general search engine/information retrieval system as the starting point during a health information search session.
	Q:AccHW	Access a consumer health informatics website as the starting point during a health information search session.
	Q:NewQ	Issue a new query, which is usually the first query in the search session.
	Q:ModQ	Reformulate the previous query to obtain more general/specific retrieval results.
Accessing	A:SelHI	Select and access a retrieved item from a health/medical website.
	A:SelGI	Select and access a retrieved item from a general/non-health-specific website.
	A:XplorF	In the retrieved item selected, access a link to another web page that has not been visited before.
	A:AccB	Access a previously visited web page using the browser's back button, by following hyperlinks, or by tracking the history.
Evaluating	E:ExamSR	Examine the results retrieved to identify items (web pages) that contain potentially relevant health information.
	E:DisSR	Discard the results retrieved with or without examining their relevance.
	E:EvalI	Evaluate the selected item from the retrieved results or visit a web page to determine its relevance.
	E:FindQ	Search for a specific keyword on a visited webpage.
Using	U:UseHI	Assess the visited health/medical web page as a relevant source and use the information it contains to answer the questions in the search task.
	U:UseGI	Assess the visited general/non-health-specific web page as a relevant source and use the information it contains to answer the questions in the search task.

Stage	Search Activity Code	Description
Discarding	D:DisHI	Assess the visited health/medical web page as irrelevant.
	D:DisGI	Assess the visited general/non-health-specific web page as an irrelevant source.
	D:UnchkHI	Discard the selected health/medical web page without visiting and evaluating its relevance.
	D:UnchkGI	Discard the selected general/non-health-specific web page without visiting and evaluating its relevance.

To begin a health information search session, a consumer accesses a general search engine (Q:AccSE) or visits a known consumer health website (Q:AccHW). The consumer's familiarity with health topics may influence the starting points he/she selected. The subsequent activities in querying stage are submitting a new query (Q:NewQ) and reformulating a query (Q:ModQ). These activities are differentiated because the type of query (new or modify) may have reflected the consumer's information base, such as background knowledge and familiarity with the search topic. In the evaluation stage, there are some differences in examining the search results (E:ExamSR) and in evaluating an individual item (E:EvalI), thus both activity types are included. When examining the search result, the consumer may not select a specific item/document from the results retrieved (E:DisSR). The evaluation stage also involves finding the query keyword (E:FindQ) because it may indicate an advanced evaluation strategy or difficulty understanding the content. In the accessing stage, selecting an item from the results retrieved is included because it reflects the consumer's ability to locate a potentially relevant source.

Next, the item selection is divided into two types: selecting a result from a health/medical specific website (A:SelHI) and selecting a result from a general website (A:SelGI), considering that the familiarity with the search topic may influence the domain type selected. The next codes, i.e., exploring forward (A:XplorF) and accessing backward (A:AccB), are treated as different activity types because the direction of accessing has different meanings in the search process

[35]. The next stages, i.e., using and discarding, were included to study the consumer's behavior when assessing the web pages they visited, and to determine the efficiency and the success/failure rate of the overall search process.

The type of web page selected and the type of relevant / irrelevant web pages are important in health information search because the source of web page determines the quality of health information. Consumer is required to examine carefully the provider who is responsible for health content in a selected / visited website. Health information must comply rigorous regulation before can be published online. Only trustworthy and reliable health information can be used as the source for learning a specific disease, educating oneself with healthy lifestyle, making a health-related decision, and any other health related needs.

3.3.2 Query Formulation and Search Result Interaction

The second factor characterizing the consumer's behavior in health information search is the consumer's query formulation. This factor includes the following features:

1. Query formulation pattern, this attribute refers to the pattern of all issued health query in a search session, e.g., how the consumer creates the initial health query and how the corresponding query is modified in the subsequent search iteration.
2. Query statistics: the total query in a search session, the average query length, and the frequency of each query type.
3. The query type, this attribute refers to the source and the type of query keywords. A consumer constructs the keywords in a health query by recalling his/her own knowledge and experience or by quoting the terminology from health references. This thesis classifies the terminology in a health query into lay (consumer-friendly) and advanced health query based on its average frequency score in Consumer Health Vocabularies [6].

Expertise or familiarity with the search topic also influences consumer's interaction with the search result. Familiar, somewhat familiar, and unfamiliar consumers exhibit different behavior in locating the potential relevant web pages, accessing the selected web pages, and assessing the relevant health information.

Chapter 4

Experimental Design Supporting Individual Health Topic Familiarity in Health Information Search

4.1 Instrument

The instrument used for data collection comprised of a health terminology familiarity questionnaire and a set of health information search tasks. Both instruments considered similar health topics, i.e., skin allergy and its main treatments, cardiovascular disease, a common medical test (urinalysis), and cholesterol problems. The health topics selected for this study were based on the common health topics discussed in Yahoo Health [56] to ensure that the experiment reflected real-life health information searches.

4.1.1 Health Terminology Familiarity Questionnaire

The familiarity questionnaire was modeled on the basis of the Familiarity of Sample Terms Questionnaire [14], the CHV Health Vocabulary Questionnaire [15], and the Test of Functional Health Literacy in Adults (TOFHLA) [57]. The questionnaire facilitated the rapid estimation of the familiarity of participants with predefined health topics. Figure 4.1 shows examples of the questions in the questionnaire and the entire questionnaire is available in Appendix A.

The questionnaire comprised three sections, each of which addressed the same four health topics. There were eight questions in each section. The questions with the same number in each section were equivalent. Section 1 estimated recognition at the surface level, while Sections 2 and 3 estimated the conceptual understandings of consumer-friendly terminology and the conceptual understandings of advanced health terminology, respectively. Each correct answer in the questionnaire was awarded 0.15 points for Section 1 and 0.175 points for

Sections 2 and 3. The familiarity category was assigned to each health topic for each participant based on the total points awarded for the health topic (six questions). The labeling method employed in this thesis refers to category of health topic familiarity in Section 3.1, as follows:

1. Category L1 (unfamiliar) was assigned to a participant with total points ≤ 0.3.
2. Category L2 (somewhat familiar) was assigned to a participant with total points > 0.3 and ≤ 0.65.
3. Category L3 (familiar) was assigned to a participant with totals points > 0.65.

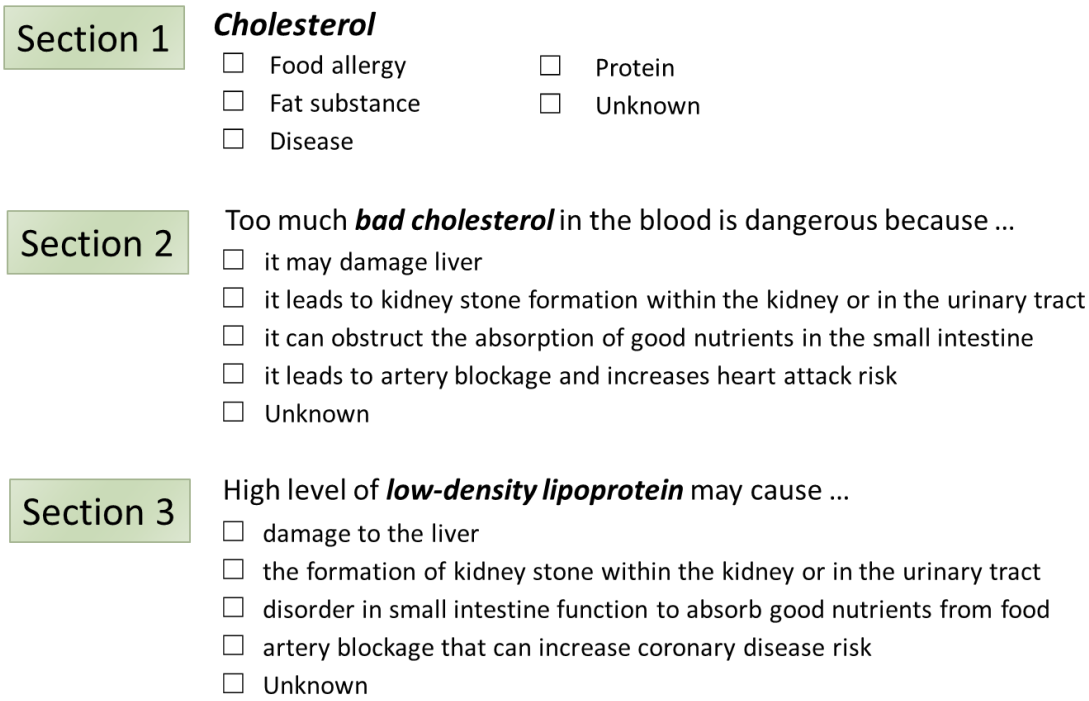


Figure 4.1 Examples of the questions included in the health terminology familiarity questionnaire

4.1.2 Health Search Task

The second instrument, Health Information Search Tasks, aimed to determine their search behaviors. The search tasks comprised four separate tasks, each of which simulated one of the predefined health topics found in the questionnaire. A short scenario was added to each task to provide a context as presented in Table 4.1.

Table 4.1 Health search tasks

<table>
<tr><th>Task ID</th><th>Task Description</th></tr>
<tr><td>Task 1</td><td>During the past six days, your skin has been very itchy and dry, particularly in your arm, wrist, and leg areas. You also noticed the appearance of rashes and redness on your itchy skin. You want to find out what might happen to your skin and how to treat it.</td></tr>
<tr><td>Task 2</td><td>In a first aid training course, your instructor emphasizes that lay people need to understand sudden cardiac arrest (SCA). SCA is often equated incorrectly with a heart attack, but SCA victims can survive if they receive treatment within 3–5 min after they collapse. You want to know:
1. The difference between a heart attack and a SCA.
2. How a lay person can help a victim when a suspected SCA incident happens in a public area.</td></tr>
<tr><td>Task 3</td><td>Every year your institution holds a mandatory general medical check-up. One of the medical tests is urinalysis. You usually receive the results about three weeks after the test.

You want to know the purpose of each parameter (why each parameter is tested) in the sample below and the meaning of the results (normal or abnormal).

Parameter — Result
Specific gravity — 1.030 (reference interval: 1.002–1.030)
pH — 4.9 (reference interval: 4.6–7.5)
Protein — Negative (reference interval: negative)
Glucose — 100 mg/dL (reference interval: negative)</td></tr>
<tr><td>Task 4</td><td>Your doctor prescribed simvastatin and instructed you not to consume the medicine with grapefruit juice. You want to know the purpose of simvastatin and why it should not be consumed with grapefruit juice.</td></tr>
</table>

The answers for the entire health questions in this search task can be found in:

- General consumer health informatics websites, such as MedlinePlus (http://www.nlm.nih.gov/medlineplus/), WebMD (http://www.webmd.com), and MayoClinic (http://www.mayoclinic.org/),

- Health community / medical association websites, such as Heart.org (http://www.heart.org/HEARTORG/, and Sudden Cardiac Arrest (SCA) Foundation website (http://www.sca-aware.org/).
- Medical journal listed in PubMed (http://www.ncbi.nlm.nih.gov/pubmed).

The participants were expected to answer the questions easily. They can choose the correct answer from any sources according to their preference or familiarity.

4.2 Data Collection Procedure and Data Analysis

In this study, the participants were observed in an experimental setting. A controllable environment and standardized health information search tasks are required to examine the effects of different parameters on the behaviors of participants. The data were collected in a private laboratory. Upon arrival, the participant was welcomed and given a brief introduction to the purpose of this study, instructions about how to complete the questionnaire, and the procedure of the search tasks. The participants were also asked to review a consent form. Each participant performed the data collection process in the following order.

1. Demographic profile survey: The participant provided demographic information and details of their experiences with health information search on the Internet.
2. Health terminology familiarity questionnaire: The participant completed the questionnaire from Sections 1 to 3 in chronological order. If the participant had never heard of the term used in the question, the participant was requested not to guess the answer and instead they were asked to select the option "Unknown."
3. Health information search task session: The participant was asked to complete the search tasks one by one. The participant was free to use any search engines or health information retrieval systems, to access any relevant websites, and to search at their own speed. Videos of all the search sessions were recorded using Camstudio screen and audio recording software [58].

After completing each task, the participant provided comments about the search topic and the search session.

The data collected from the participants comprised demographic data, responses to the familiarity questionnaire, and video recordings of the health information search sessions. The demographic data were used to capture the general characteristics of the participants. The responses to the familiarity questionnaire were utilized to label the familiarity of participants with the predefined health topics. The participants were categorized into three familiarity categories (L1, L2, and L3). The search outcome (participant's answer) from health information search task session was measured as relevant (correct) or not relevant to the question. Subsequently, the video data that contained the finding of relevant answer was transcribed and analyzed.

4.3 Demographic Profile of the Participants

A convenience sample of 40 participants was recruited from several departments of two universities (Table 1). The participants were undergraduate students, exchange students, graduate students, and researchers from the Engineering, Material Physics, Applied Physics, Biotechnology, Information and Physical Sciences, and Computer Science departments. University students and researchers were selected as the target participants because they were part of general consumers in the real world and the data collection process required a certain degree of cognitive effort.

The criteria for the recruitment of participants were non-medical professionals, the ability to read and write in English, the ability to use the computer and Internet, and age ≥18 years. The participants come from Japan, Indonesia, Thailand, Philippines, United States of America, Germany, and Spain. The selected samples have covered three categories of familiarity (unfamiliar, somewhat familiar, and familiar) in four different health topics (skin allergy and its main treatments, cardiovascular disease, urinalysis medical test, and cholesterol problems) that were used in this study. All participants had experience in health information search on the Internet before the study was conducted.

Table 4.2 Demographic profile of the participants

Demographic Profile	Categories	*N*	%
Gender	Male	24	60
	Female	16	40
Age	18–25 years	28	70
	26–35 years	12	30
	36–45 years	0	0
	> 45 years	0	0
Native language	English	15	38
	Non-English	25	62
Education	High School	0	0
	Bachelor's degree	22	55
	Graduate degree	18	45
Health information seeking experience	Frequently on daily / weekly basis	8	20
	Occasionally on monthly basis	7	17
	Yearly or less than five times ever	5	12
	As the need arises	20	50
	Never	0	0

4.4 Health Topic Familiarity Questionnaire Result

Table 4.3 shows the result of familiarity labeling for each health topic based on the responses to the familiarity questionnaire. Each participant produced four data instances, i.e., one for each health topic; hence the overall data collection resulted in 160 instances. According to this result, a participant could have different familiarity category labels for different health topics. For example, a participant was highly familiar with topics 2 and 4, but unfamiliar with topics 1 and 3. Another participant was familiar with topic 1, somewhat familiar with topic 3, and unfamiliar with topic 2 and 4.

Table 4.3 The familiarity questionnaire result

No.	Health Topic	L1	L2	L3	Number of participants
1	Skin allergy and main medications	14	9	17	40
2	Cardiovascular disease	12	19	9	40
3	Common medical test (urinalysis)	17	11	12	40
4	Cholesterol problems	18	12	10	40
	Total instances	61	51	48	160

This page is intentionally left blank.

Chapter 5

The Effects of Health Topic Familiarity on Health Information Search Behavior

This chapter discusses the importance of health topic familiarity in health information search process to answer the first research question. It includes the detail explanation of the method, the result, and the analysis of the result.

5.1 Method of Examining the Effects of Health Topic Familiarity on Search Behavior

As the initial effort toward the improvement of the health information search process, we studied the importance of health topic familiarity by examining its effects on the consumer's search behavior. For this purpose, we developed a method to examine the participant's search behavior, as presented in Figure 5.1.

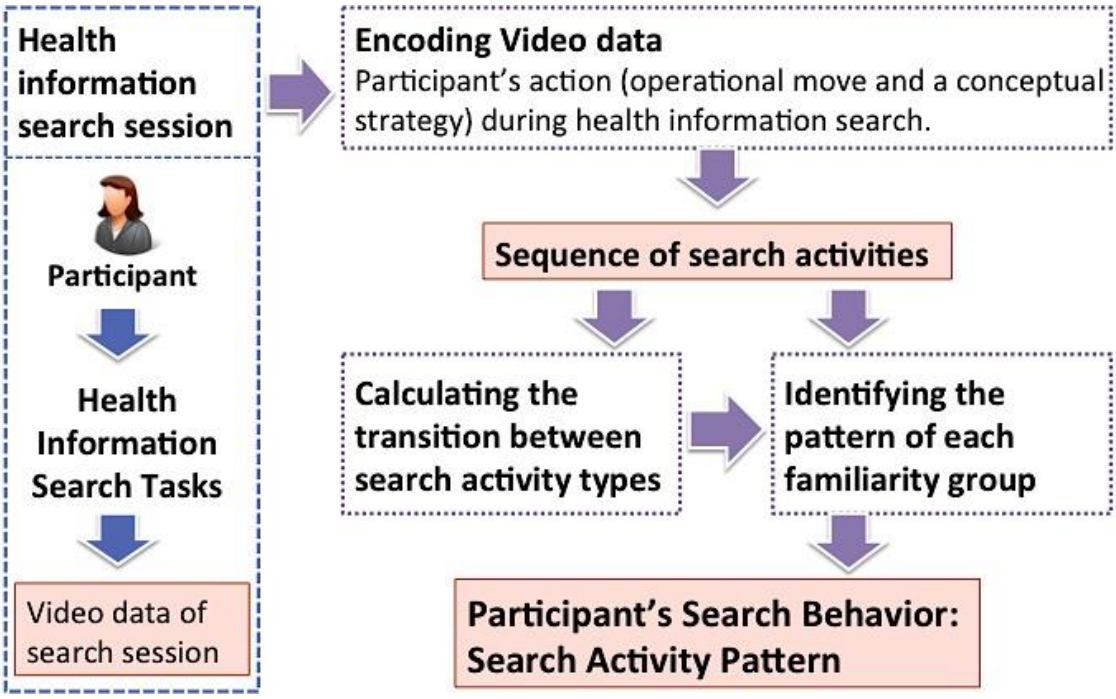


Figure 5.1 The method of examining participant's search behavior in health information search

The method comprised of three steps, i.e., modeling the search session video data into sequence of search activities, calculating the transition between search activity types, and identifying the common search activity pattern for each familiarity group. The following subsections describe in detail each step in the method.

5.1.1 Modeling the Search Session

Each search session, which included the finding of correct answer, was encoded as a sequence of search activities, using the coding scheme in Table 3.1. For example, a search session of Task 1 from a participant exhibited twenty search activities, as follows:

1. The participant started the search session by accessing a general search engine, submitting the first query, and examining the results retrieved (Q:AccSE–Q:NewQ–E:ExamSR).
2. The participant selected an item from a health website and an item from a non-health-specific website (A:SelHI–A:SelGI).
3. The participant evaluated the first item selected and assessed it as a relevant source (E:EvalI–U:UseHI).
4. Next, the participant evaluated the second item selected and assessed it as a non-relevant source (E:EvalI–D:DisGI).
5. The participant continued the search process by modifying the previous query and examining the results retrieved (Q:ModQ-E:ExamSR).
6. The participant selected three items from health websites (A:SelHI– A:SelHI-A:SelHI).
7. The participant evaluated the first item selected and assessed it as a non-relevant source (E:EvalI–D:DisHI).
8. Next, the participant evaluated the second item selected and assessed it as a non-relevant source (E:EvalI–D:DisHI).
9. Lastly, the participant evaluated the third item selected and assessed it as a relevant source (E:EvalI–U:UseHI).

5.1.2 Calculating the Transition Frequency between Search Activity Types

To examine how the participants progressed during their search process, the next step involved calculating the transition frequencies and the probabilities between the states of all possible search activity types. Given a collection of mutually exclusive states (such as the search activity types in this study), the first-order transition probability in a Markov model gives the probability of moving from one state to another [49]. In this study, the transition probabilities were calculated on the basis of a first order Markov model.

After calculating the transition frequency and probability for each familiarity group, the chi-squared test was performed at a significance level of α = 0.01 to verify the differences in the search activity transitions between familiarity groups. The null hypothesis was that there was no difference in the first order state transition probability matrices between familiarity groups. The test followed the procedure reported by Chen and Cooper [49], as follows:

1. Let A and B be the two samples that need to be compared. A transition frequency matrix for sample A is defined as f_{ij}^{A} *(i,j = 1, 2, …, K)*, where f_{ij}^{A} is the number of transitions from state *i* to state *j*, and *K* is the number of states in the state space.
2. If sample B is similar to sample A, then f_{ij}^{B} should be close to the expected number of transitions from state *i* to state *j* in B, as follows.

$$E\left(f_{ij}^{B}\right) = \sum_{l=1}^{K} f_{il}^{B} * \frac{f_{ij}^{A}}{\sum_{l=1}^{K} f_{il}^{A}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

3. In this case, the value *C* obtained from the following equation:

$$C = \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{[f_{ij}^{B} - E(f_{ij}^{B})]^2}{E(f_{ij}^{B})} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

will approximate a chi-square distribution with degrees of freedom: $K^2 - N_1 - N_2$, where N_1 is the number of actual states in B and N_2 is the number of impossible transitions in B. The null hypothesis that there is no difference between transition probability matrices A and B is accepted if C is less than the critical value of $C_a^{K^2 - N_1 - N_2}$ at a significance level of α = 0.01.

5.1.3 Identifying Search Activity Patterns

To better understand and characterize the search behaviors of different familiarity groups, the next step in the data analysis process was to discover common search activity patterns using the following method:

1. Building an n-gram language model of the sequence of search activities performed by participants based on the dataset.

   An n-gram model is a probabilistic language model, which is used to predict the next word from a sequence of word [35]. When estimating an n-gram model, it is normally assumed that the sequence histories of words depend only on the local prior context (Markov model assumption) because of the large number of parameters involved [46]. To build n-gram language model, we utilized the SRI Language Modeling toolkit [59] and four datasets (L1, L2, L3, and the data for all participants) with the Witten-Bell discounting strategy [60]. Each dataset was divided into 80% training data and 20% test data. The n-gram language models were built using the training data with various sequences, i.e., 2-grams to 7-grams.

2. Evaluating the perplexity of the computed language models to specify the number of search activities in a sequence that best represented the search activity pattern.

   The perplexity of a language model represents the geometric average branching factor of the language according to the model and is used widely to measure the quality of a model (lower perplexity tend to have lower word-error rates) [61]. The perplexity *PP(pM)* of a language model *pM (next word w|history h)* on a test set $T = \{w_1, \ldots, w_t\}$ is computed using the following equation:

$$PP_T(p_M) = \frac{1}{(\prod_{i=1}^{t} p_M(w_i|w_1 \ldots w_{i-1}))^{\frac{1}{t}}} \qquad \ldots\ldots\ldots\ldots\ldots (3)$$

   This metric was used because the computed language models contained similar vocabularies (i.e., the search activity types). The number of search activities in a sequence was represented by the n-gram sequence with the lowest perplexity.

3. Applying the selected *n*-gram model to the sequence of search activities in the datasets to identify common search activity patterns.

5.2 Result

The experiment in this chapter produced three main results, i.e., frequency of search activities, transition between search activities, and most frequently pattern of search activities sequence applied in each familiarity group.

5.2.1 Frequency of Search Activities

All of the search sessions performed by the 40 participants contained the finding of the correct answer to the questions in Health Search Task. Thus, all of the video data were transcribed and produced 4595 search activities. Figure 5.2 and Table 5.1 show the breakdown of the search activities in each familiarity groups. The number of search activities in a health information search session varied from 6 to 221. On an average, a participant performed 28.7 search activities during one health information search session (SD = 23.27).

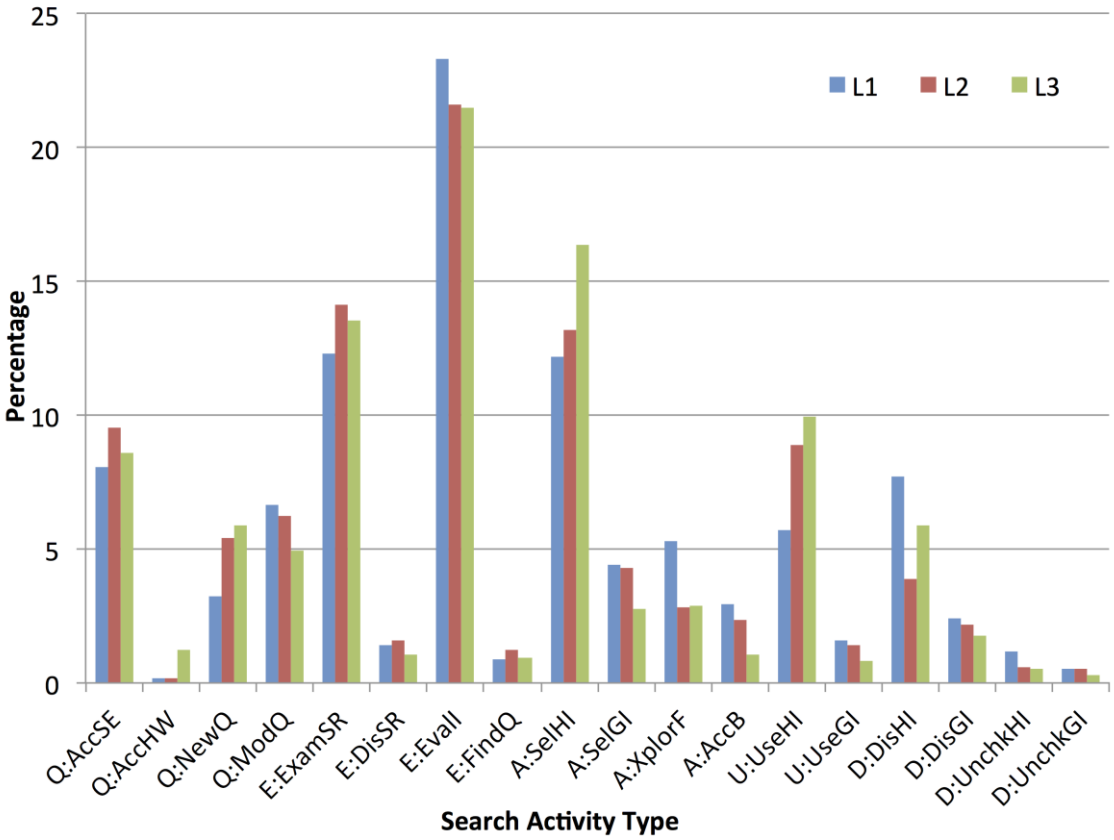


Figure 5.2 Percentage of the search activity types in all familiarity groups

Table 5.1 Frequency and proportion of search activity type

Search Activity Type	Frequency			Proportion (%)		
	L1	L2	L3	L1	L2	L3
Q-AccSE	196	115	83	8.09	9.55	8.58
Q-AccHW	4	2	12	0.17	0.17	1.24
Q-NewQ	78	65	57	3.22	5.40	5.89
Q-ModQ	162	75	48	6.68	6.23	4.96
E-ExamSR	298	170	131	12.29	14.12	13.55
E-DisSR	34	19	10	1.40	1.58	1.03
E-EvalI	565	260	208	23.31	21.59	21.51
E-FindQ	21	15	9	0.87	1.25	0.93
A-SelHI	295	159	158	12.17	13.21	16.34
A-SelGI	107	52	27	4.41	4.32	2.79
A-AccF	129	34	28	5.32	2.82	2.90
A-AccB	72	28	10	2.97	2.33	1.03
U-UseHI	139	107	96	5.73	8.89	9.93
U-UseGI	38	17	8	1.57	1.41	0.83
D-DisHI	187	47	57	7.71	3.90	5.89
D-DisGI	59	26	17	2.43	2.16	1.76
D-UnchkHI	28	7	5	1.16	0.58	0.52
D-UnchkGI	12	6	3	0.50	0.50	0.31
Total	**2424**	**1204**	**967**	**100.00**	**100.00**	**100.00**

The most frequent search activity in all the familiarity groups was evaluating a selected item from the results retrieved (E:EvalI). This search activity accounted for 562 out of 2424 (23.31%) activities in group L1, 260 out of 1204 (21.59%) in group L2, and 208 out of 967 (21.51%) in group L3. The second, third, and fourth most frequent search activities in groups L1 and L2 were examining the results retrieved (E:ExamSR), selecting a health-related item from the results retrieved (A:SelHI), and accessing a general search engine (Q:AccSE), which together

comprised 32.55% (789/2424) of the activities by group L1 and 36.88% (444/1204) by group L2. In contrast to these groups, A:SelHI, E:ExamSR, and U:UseHI were the second, third, and fourth most frequent search activities among participants in group L3, which together represented 39.81% (385/967) of the total. The fifth most frequent search activities were discarding the selected health-related website (D:DisHI), U:UseHI, and Q:AccSE for groups L1, L2, and L3, respectively.

All of the groups exhibited the same pattern when accessing the results retrieved. Participants were more likely to access health/medical websites than general domain websites. Group L3 accessed health websites more frequently than others, 85.42% (158/185) compared with 73.4% (159/211) and 75.36% (295/402). In contrast, group L1 accessed more general domain websites (26.60%, 107/402) than group L2 (24.64%, 52/211) and group L3 (14.58%, 27/185). In terms of locating the relevant health information, the participants in all groups tended to engage in a considerable number of search activities before reaching U:UseHI or U:UseGI. The combinations of U:UseHI and U:UseGI in groups L1, L2, and L3 were 7.30% (177/2424), 10.30% (124/1024), and 10.75% (104/967), respectively. Those proportions were relatively smaller compared to querying, accessing, and evaluating search activities.

5.2.2 Transition between Search Activity Types

Table 5.2 provides most frequent transitions between search activities. The calculations yielded a total of 4435 transitions, i.e., 2363 transitions, 1153 transitions, and 919 transitions in groups L1, L2, and L3 respectively. The average numbers of transition between two search activities were 19.86 (SD = 24.70) in group L1, 14.06 (SD = 13.26) in group L2, and 11.78 (SD = 11.38) in group L3. The most frequent transitions in all groups were related to accessing a health website from the results retrieved and evaluating its relevancy. The corresponding transitions were from E:ExamSR to A:SelHI (L1 = 7.96%, L2 = 9.80%, L3 = 11.1%) and from A:SelHI to E:EvalI (L1 = 7.66%, L2 = 8.76%, L3 = 11.0%).

The third most frequent transition in the unfamiliar group (L1) was different from that in the other more familiar groups (L2 and L3). The transition in group L1 from E:EvalI to D:DisHI showed that the participants assessed the selected item as irrelevant. In contrast, the third most frequent transition in groups L2 and L3 was from E:EvalI to U:UseHI. This finding indicates that the participants in L2 and L3 were probably more successful than those in L1 at identifying potentially relevant items from the results retrieved.

Table 5.2 Top 10 frequent first order transitions for each familiarity group

No.	L1			L2			L3		
	Transition	Frequency		Transition	Frequency		Transition	Frequency	
		N	%		*N*	%		*N*	%
1	E:ExamSR–A:SelHI	188	7.96	E:ExamSR–A:SelHI	113	9.80	E:ExamSR–A:SelHI	102	11.1
2	A:SelHI–E:EvalI	181	7.66	A:SelHI–E:EvalI	101	8.76	A:SelHI–E:EvalI	101	11.0
3	E:EvalI–D:DisHI	160	6.77	E:EvalI–U:UseHI	94	8.15	E:EvalI–U:UseHI	81	8.8
4	Q:ModQ–E:ExamSR	158	6.69	Q:ModQ–E:ExamSR	75	6.50	Q:NewQ–E:ExamSR	56	6.1
5	Q:AccSE–Q:ModQ	121	5.12	Q:NewQ–E:ExamSR	64	5.55	E:EvalI–D:DisHI	51	5.6
6	E:EvalI–U:UseHI	120	5.08	Q:AccSE–Q:NewQ	63	5.46	Q:AccSE–Q:NewQ	48	5.2
7	A:XplorF–E:EvalI	91	3.85	Q:AccSE–Q:ModQ	52	4.51	A:SelHI–A:SelHI	44	4.8
8	E:EvalI–A:XplorF	88	3.72	A:SelHI–A:SelHI	40	3.47	Q:ModQ–E:ExamSR	44	4.8
9	Q:AccSE–Q:NewQ	75	3.17	A:SelGI–E:EvalI	39	3.38	Q:AccSE–Q:ModQ	35	3.8
10	Q:NewQ–E:ExamSR	75	3.17	E:EvalI–D:DisHI	36	3.12	A:XplorF–E:EvalI	24	2.6
Total		1257	53.20		677	58.72		586	63.8

During the querying stage, group L3 had different search activities compared with the other less familiar groups (L1 and L2). The most frequent transition related to the querying stage was from Q:NewQ to E:ExamSR in group L3 and from Q:ModQ to E:ExamSR in groups L1 and L2. This shows that the L3 participants probably relied on their first query to discover relevant results. Group L3 also performed less query modifications than the other groups.

The complete transition probability matrix in group L1, L2, and L3 is available in Appendix B.

5.2.3 Testing the Differences in the Search Activities between Familiarity Groups

Table 5.3 shows the result of the chi-squared test described in the Subsection Calculating the Transition Frequency between Search Activity Types. According to the results, the null hypothesis was rejected in all cases; hence, the three familiarity groups exhibited distinct search activity patterns.

Table 5.3 Results obtained after testing the differences between the familiarity groups ($P < .001$)

Familiarity Group	L1	L2	L3
L1	-	$K^2 = 324$ $N_1 = 18$ $N_2 = 242$ $df = 64$ ($\chi^2 = 104.716$) $C = 5084.883$	$K^2 = 324$ $N_1 = 18$ $N_2 = 246$ $df = 60$ ($\chi^2 = 99.607$) $C = 6021.407$
L2	$K^2 = 324$ $N_1 = 18$ $N_2 = 204$ $df = 102$ ($\chi^2 = 151.884$) $C = 2211.996$	-	$K^2 = 324$ $N_1 = 18$ $N_2 = 246$ $df = 60$ ($\chi^2 = 99.607$) $C = 2809.463$
L3	$K^2 = 324$ $N_1 = 18$ $N_2 = 204$ $df = 102$ ($\chi^2 = 151.884$) $C = 1787.706$	$K^2 = 324$ $N_1 = 18$ $N_2 = 242$ $df = 62$ ($\chi^2 = 102.166$) $C = 1651.765$	-

K is the number of state in the state spaces of the corresponding row of the table
N_1 is the number of actual states in the corresponding column of the table
N_2 is the number of impossible transitions in the corresponding column of the table
df is obtained from K^2-N_1-N_2
C is the Chi-square score obtained from Equation (2)

5.2.4 Most Frequently Pattern of Search Activities Sequence Applied in Each Familiarity Group

According to the perplexity evaluations of all the language models for all the datasets in Figure 5.3, 5-gram language models had the lowest perplexity values for the four test datasets. Thus, we used *5-gram sequences* to identify common search activity patterns in each familiarity group. The numbers of observed 5-gram sequences in groups L1, L2, and L3 were 940, 444, and 359, respectively. There were large numbers of 5-gram sequences in each group; so only the 20 most frequent sequences were examined (the complete list of 5-gram sequences is available in Appendix C). Above this level, the frequencies of the sequences were too low to represent the search activity patterns in a familiarity group.

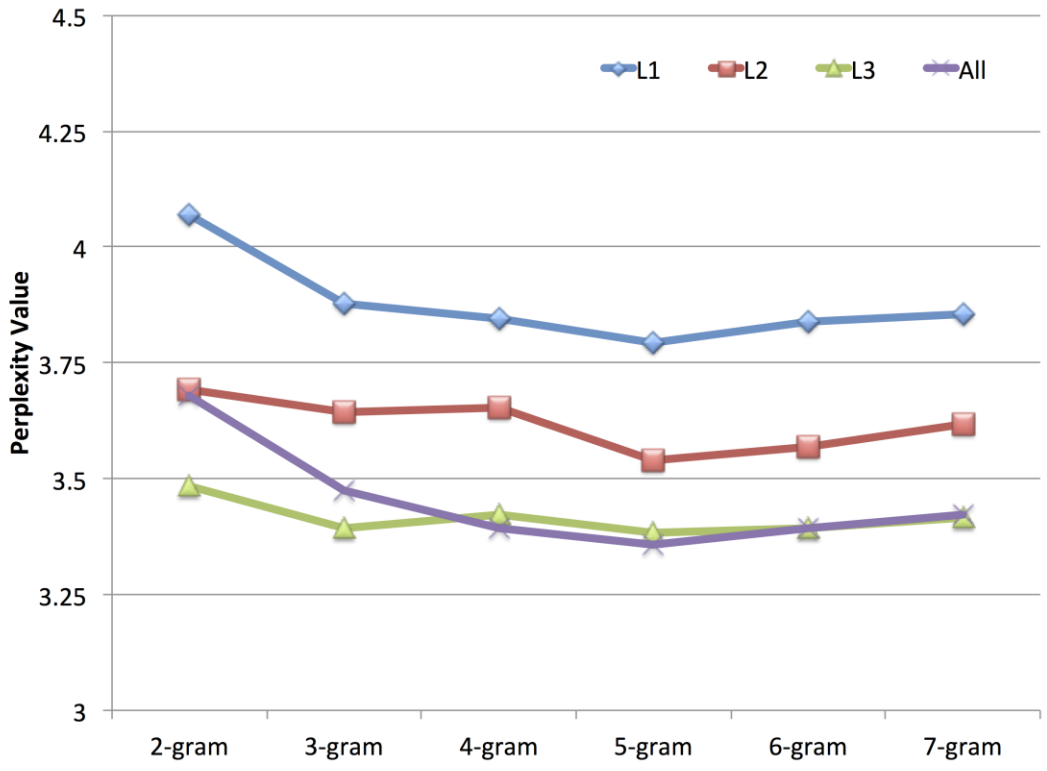


Figure 5.3 Perplexity values for L1, L2, L3, and all the test data using different n-gram models

To compare the search behavior between familiarity groups, we used four activity categories from the health information search process based on the top 20 most frequent patterns as the following:

1. Category 1: Accessing a search engine (general search engine or consumer health website), issuing a new or modified query, and accessing and evaluating an item from a health website.
2. Category 2: Accessing a search engine, issuing a query, and accessing multiple items from health websites.
3. Category 3: Accessing, evaluating an item from a health website, and assessing the relevancy.
4. Category 4: Continuing the search process after finding a relevant item.

The comparison of frequent search activity patterns in all familiarity groups is presented in Table 5.4.

Table 5.4 Comparison of frequent activity patterns

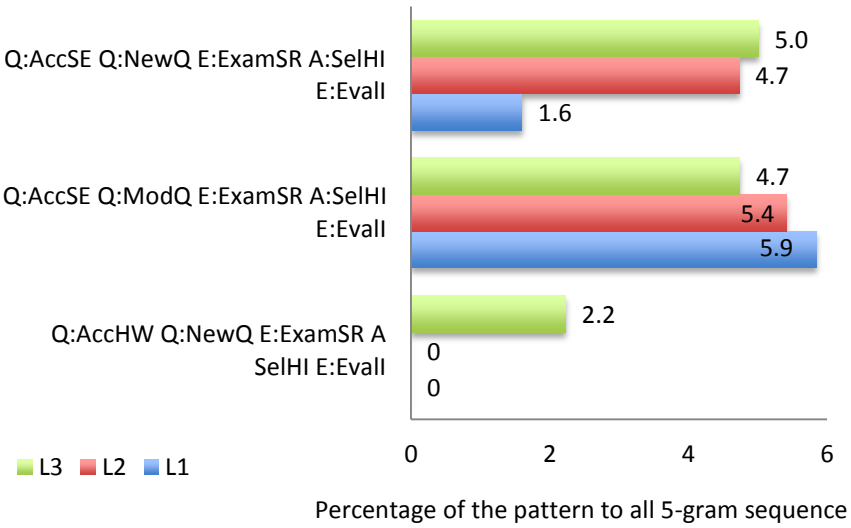
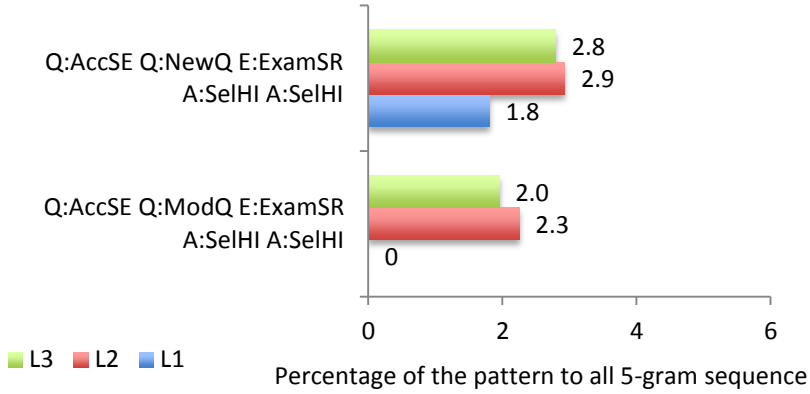
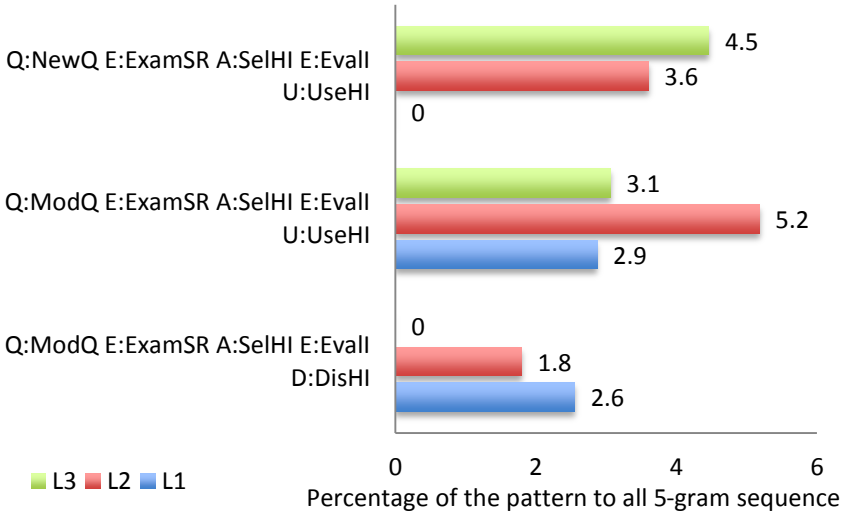
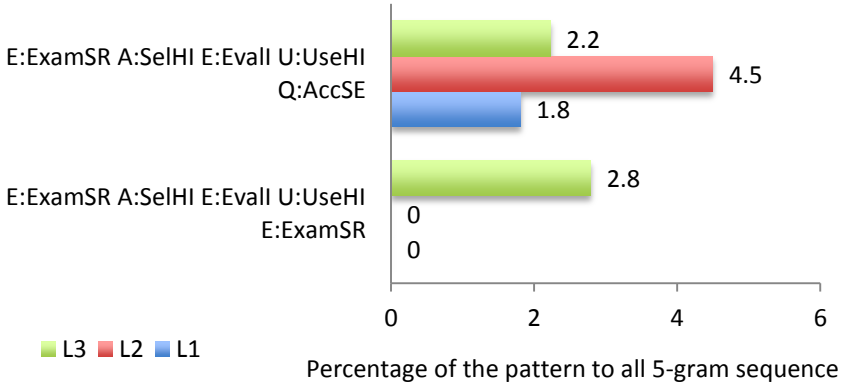
Activities during a health search session	Comparison of frequent activity patterns[a]
Category 1: Issuing a query, examining the results retrieved, accessing, and evaluating an item from a health website	

Figure 5.4 Comparison of frequent activity patterns in Category 1

Activities during a health search session	Comparison of frequent activity patterns[a]
Category 2: Issuing a query, examining the results retrieved, accessing multiple health websites	 **Figure 5.5 Comparison of frequent activity patterns in Category 2**
Category 3: Accessing, evaluating an item from a health website, and assessing the relevancy	 **Figure 5.6 Comparison of frequent activity patterns in Category 3**
Category 4: Continuing the search process after finding a relevant item	 **Figure 5.7 Comparison of frequent activity patterns in Category 4**

[a] 0 indicates that a pattern is not among the top 20 most frequent patterns

Group L1 comprised participants who were not familiar with the health topic search task. The most frequent pattern in group L1 was submitting a modified query to a general search engine, followed by accessing a health-related website from the search results, and immediately evaluating the relevancy of the selected result (Q:AccSE–Q:ModQ–E:ExamSR–A:SelHI–E:EvalI), which accounted for 5.85% of all the 5-gram patterns. In locating the potentially relevant search results, this group was accessed more non-relevant results than relevant results.

As shown in Category 3 of Table 5.4, the proportion of D:DisHI assessments was larger than that of U:UseHI assessments. In total, 10/20 of the most frequent patterns contained D:DisHI (see Appendix C), which accounted for 23.3% of all the 5-gram patterns in group L1. In contrast, only 5/20 of the most frequent patterns included U:UseHI assessments, which comprised 11.5% of all the 5-gram patterns.

In group L2, all of the queries in the top 20 most common patterns were submitted to a general search engine. The proportion that issued a modified query was higher than that issuing a new query. The identification of the potentially relevant search results showed that participants in this group were likely to be more successful than those in group L1, as demonstrated by the higher proportion of U:UseHI assessments than D:DisHI assessments. The participants in group 2 created a new search after finding a relevant information source.

The final group, L3, comprised the most knowledgeable searchers. The proportion that issued a new query was higher than that issuing a modified query. Unlike the other groups, the participants in group L3 also accessed consumer health websites to search for health information. Two strategies were performed by group L3 when accessing the search results: accessing a single item from a health website and evaluating it immediately, or accessing multiple items from health websites and evaluating the items one by one. When identifying potentially relevant search results, group L3 found more relevant items in the results retrieved from the first query compared with the results retrieved using the modified query. The

participants also continued their search process by creating a new search and reexamining the previous results retrieved.

5.3 Analysis and Discussion

The health information search behavior was characterized as a sequence of search activities in this study. The frequencies of the search activities showed that the participants devoted substantial efforts during the evaluating stage, where they examined the results retrieved and evaluated the relevancy of the item selected. The participants also performed frequent search activities during the accessing and querying stages. Although the use of selected information (U:UseHI and U:UseGI) is the main goal of information search, the total proportions of these search activities were smaller than the search activities performed in the evaluating, accessing, or querying stages. These findings indicate that health information search remains difficult for most consumers.

According to the experiment results, the participants with different levels of familiarity performed a unique search behavior, as summarized in Table 5.5. The first effect of health topic familiarity was observed in the querying stage. The participants in the lower familiarity groups submitted more queries than the participants in the higher familiarity group. The average numbers of query submissions during a health search session were 7.2, 5.0, and 4.2 in groups L1, L2, and L3, respectively. The series of query submissions reflected the searcher's progress in understanding the searched topic. The participants with less familiarity submitted more queries because they needed to increase their understanding of the search topic before they could locate relevant information. A number of participants in group L1 started the search process by searching for definitions of the health terms that appeared in the searching task. Examples of this type of query are "what is rash," "urinalysis definition," "what is SCA," "special gravity in urine?," and "what is simvastatin." This finding is different with other studies in general web-based search process [62, 63]. Liu et al. in their study reported that no differences in the number of queries issued were found between users with different levels of topic knowledge [62], while Zhang et al. stated in

their study that high level domain knowledge group issued more queries than the low level group [63]. In term of the average query length, there was no distinguishable pattern between less familiar and more familiar groups. This finding is also different with previous studies in [30, 37], which suggested expert users issued longer and more complex queries than novice users.

Another interesting finding is how the familiarity affected the selection of the relevant source (web pages). Less familiar participants were likely to choose easier content, while more familiar participants tended to use more difficult content. We measured the difficulty of the source by its readability score using Simple Measure of Gobbledygook (SMOG) formula [54].

The next effect was detected when locating relevant health information, which was estimated on the basis of the search efficiency. The search efficiency compared the proportion that used the information (U:UseHI and U:UseGI) against the number of items accessed (A:SelHI, A:SelGI, A:XplorF, and A:AccB). Group L3 achieved the best performance with a search efficiency of 46.6%, compared with 45.5% and 29.4% for groups L2 and L1, respectively. This result agreed with the frequencies of search activities in each familiarity group. Group L1 accessed more irrelevant items than relevant ones, whereas groups L2 and L3 did the opposite. This finding is in contrast to previous study, which reported that the search effectiveness remained the same for all participants in high and low levels of domain knowledge [63].

The patterns exhibited in each group also illustrated the effect of the level of health topic familiarity on search behaviors. The frequent patterns in group L1 showed that these participants were likely to experience difficulties during their health information search sessions, as demonstrated in the much higher percentage of issuing modified queries than issuing new queries and in identifying the potentially relevant search results. The participants found relevant information more often using the results retrieved with the modified query than the first query. The common strategies employed when the participants encountered search

problems were querying followed by single accessing and evaluating (… D:DisHI–Q:AccSE–Q:ModQ …), or iterative accessing and evaluating (… D:DisHI–E:ExamSR–A:SelHI …).

In group L2, the most frequent pattern was issuing a modified query, accessing a health website, and evaluating the selected item immediately. Group L2 also discovered relevant items more often using the results retrieved with the modified query rather than the first query, but they exhibited greater search efficiency compared with group L1. When examining the results retrieved, group L2 performed single accessing and the evaluation of selected items, or multiple accessing followed by evaluating the selected items one by one. Another frequent pattern in group L2 was the transition from U:UseHI to Q:AccSE. This pattern indicates that the participants attempted to continue health information searches after they found relevant health information. The aim of these further searches was either to verify the accuracy of the health information they discovered, or to search for another related health topic during the search task.

The most common patterns in group L3 were related to query submission and single selection, and the evaluation of a health web page. The participants in group L3 employed more varied keywords in their queries than the other groups. A frequent pattern in this group was accessing a known consumer health information website directly to start a health search session and search for health information (known item strategy). Several participants also referred to Pubmed articles to answer the questions in the search tasks, e.g., in Task 4 (the interaction between simvastatin and grapefruit juice). Another highly frequent pattern in group L3 was Q:AccSE–Q:NewQ–E:ExamSR–E:EvalI–U:UseHI, which represents a successful search when locating the relevant health information at the first attempt (first query submission and first item selection). A number of participants in group L3 continued the search process after they discovered relevant health information by issuing a modified query, or by reexamining the previous results retrieved.

Table 5.5 Summary of the findings from the experiment in Chapter 5

Familiarity Group	Characteristic frequent patterns
L1	• More likely to reformulate the query: the proportion of frequent patterns that contained a modified query (Q:ModQ) was higher than that containing the first query (Q:NewQ). • More likely to encounter difficulty during the search process, e.g., they frequently accessed irrelevant websites and had a low search efficiency. • Discovery of relevant web pages (information source) more frequently in the results were retrieved with the modified query than the first query.
L2	• More likely to reformulate the query: the proportion of frequent patterns that contained a modified query (Q:ModQ) was higher than that containing the first query (Q:NewQ). • Discovery of relevant web pages (information source) more frequently in the results were retrieved with the modified query than the first query. • Achievement of better search efficiency than group L1. • Continuation of the search process after discovering relevant web pages by issuing another query.
L3	• Access of consumer health information websites directly to start the search session. • Discovery of relevant web pages (information source) more frequently in the results were retrieved with the first query than the modified query. • Continuation of the search process by issuing another query or by reexamining the results retrieved.

This page is intentionally left blank.

Chapter 6

Prediction Model of Health Topic Familiarity based on Health Information Search Behavior

According to the result in Chapter 5 and the finding in [64], each group of consumers (unfamiliar, somewhat familiar, and familiar groups) exhibited unique search behavior. A health information search system can use this knowledge to identify the health topic familiarity for every consumer by analyzing his/her search behavior. For example, if the system detects multiple query reformulations pattern without any activities on the retrieved result, the system would consider that the consumer, who is currently using the system, experiences difficulty because of his/her unfamiliarity with the search topic. Then, the system creates a personalized model for each consumer and delivers relevant results based on consumer's familiarity.

This chapter discusses the development of prediction model of individual health topic familiarity based on health information search behavior. The explanation begins from the features extraction, the model development process, features selection, the prediction model performance, and ends with the analysis of the result.

6.1 Features

The features used to develop the familiarity prediction model were extracted from the search activity in Chapter 5 as shown in Figure 6.1. The search activity type comprised of five stages: querying, accessing, evaluating, using, and discarding. The querying stage (Q:AccSE, Q:AccHW, Q:NewQ, and Q:ModQ) and part of evaluating stage (E:ExamSR and E:DisSR) were extracted into query formulation feature set. It included query formulation pattern (how the consumer construct the first query and reformulate the query), total query, average query length, first query type, and successive query type features. The complete query features set is

shown in Table 6.1. The query type for this experiment was described in Table 6.2; hence the frequency of each query type feature was breakdown into frequency of TR query, frequency of TM query, frequency of NL query, and frequency of NM query features.

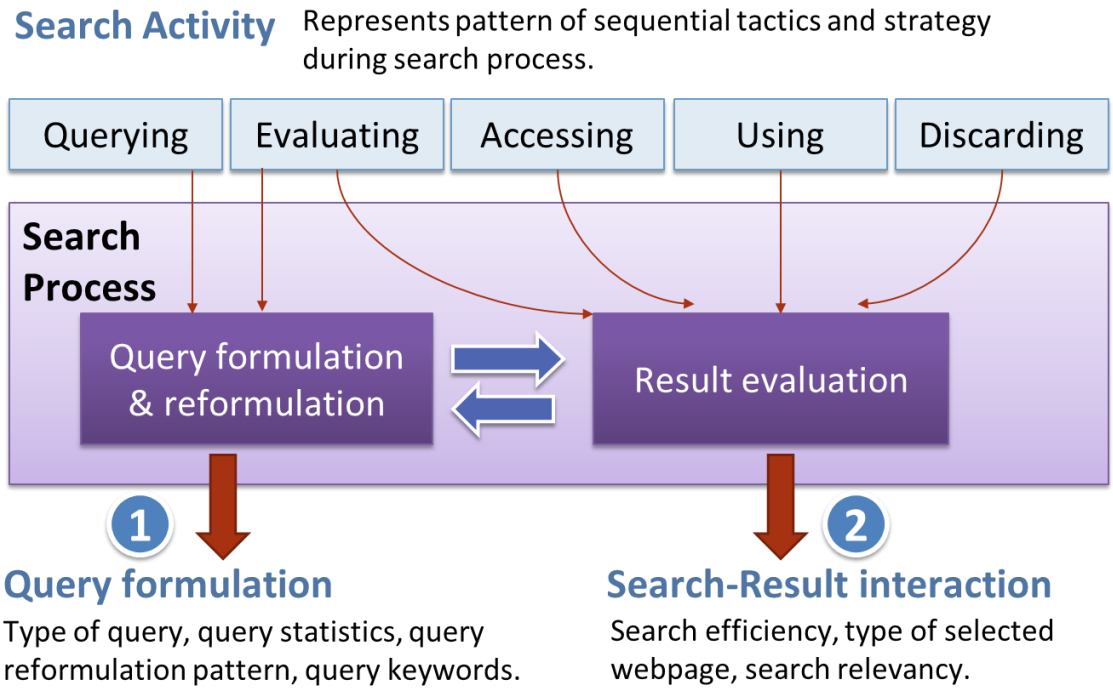


Figure 6.1 The extraction scheme from Search Activity Types to Features Sets

Table 6.1 The query formulation features

Feature	Description
Query formulation pattern	▪ The pattern of sequential query formulation and reformulation in a search session. The coding of this feature was adopted from the query reformulation pattern model of Rieh and Xie [9]: Generalization, Specification, Building Block, Parallel Movement, Dynamic, and Repetition. ▪ *No-reformulation* pattern is added to code single query submission in a search session.
Total query	The number of query submitted in a search session.
Average query length	The average number of keyword(s) in a query.

Feature	Description
First query type	The type of the first query submitted in a search session.
Successive query type	The type of the modified query submitted in a search session.
Frequency of each query type	The number of each query type submitted in a search session.

Table 6.2 Query type definition

Source of terms	Terms classification	Code	Definition and Example
Task description of Health Search Task	Repetition	TR	All terms in the query were taken from the task description in the same order. E.g., *some rashes and redness appeared on your itchy skin*.
	Modification	TM	All terms in the query were taken from the task description with modification in word's order. E.g., *skin redness dry and itchy*, *heart attack and cardiac arrest*.
New term(s)	Consumer-friendly (lay)	NL	The keyword contained at least a consumer-friendly health terminology from the participant. The average frequency score of the query is ≥ 0.5. E.g., *differences between contact dermatitis and eczema*.
	Advanced medical	NM	The keyword contained at least an advanced health / medical terminology from the participant. The average frequency score of the query is < 0.5. E.g., *low-density lipoprotein medication*, *routine and microscopy ph*.

The rest of the stages and part of evaluating stages (E:EvalI) were extracted into search result interaction set. This feature set consisted of total visited web pages, average visited web pages for each query, ratio of searching to browsing, total relevant web pages, search efficiency, ratio of relevant web pages from browsing

to from searching, average webpage dwell time, source of first relevant web page, and average reading score of selected relevant web pages features. The description of search-result interaction's feature refers to Table 6.3.

Table 6.3 The search result interaction features

Feature	Description
Total visited web pages	The total number of visited web pages in a search session.
Average visited web pages for each query	The average number of visited web pages for each query issued in a search session.
Ratio of searching to browsing	Ratio of visited web pages from searching to visited webpages from browsing in a search session. Higher number means the session is search-intensive, while lower number means exploration-intensive.
Total relevant web pages	The number of visited web page, which is assessed as a relevant source.
Search efficiency	The proportion of relevant web pages to all visited webpages in a search session.
Ratio of relevant web pages from browsing to from searching	Ratio of relevant web pages from browsing to relevant web pages from searching.
Average webpage dwell time	The average display time of visited web pages.
Source of first relevant web page	The query source of the first relevant result (FQ or MQ). FQ means the first relevant result is retrieved from the first query, while MQ means the first relevant result is retrieved from the modified /reformulated query.
Average reading score of selected relevant web pages	The average reading score of relevant web pages. The reading score is calculated using Simple Measure of Gobbledygook (SMOG) formula [54]. The formula is the preferable measure to evaluate the readability of consumer-oriented health material [55].

6.2 Features Selection

We performed the feature selection process to examine which features have strong discriminative power in classifying health topic familiarity of the consumers and to eliminate irrelevant features. This process was done in 10 fold cross validation by measuring the information gain [65, 66] of each feature with respect to the familiarity class. The information gain (as shown in the average merit value in Table 6.4) indicates the importance of the feature. The higher the average merit value, the more important the feature is.

Table 6.4 Information gain of each feature

No.	Average merit ± SD	Feature	Feature Set
1.	0.558 ± 0.021	First query type	Query formulation
2.	0.541 ± 0.028	Query reformulation pattern	Query formulation
3.	0.505 ± 0.045	Search efficiency	Search result interaction
4.	0.418 ± 0.032	Average reading score of selected relevant webpages	Search result interaction
5.	0.419 ± 0.084	Successive query type	Query formulation
6.	0.399 ± 0.017	Frequency of NM query	Query formulation
7.	0.384 ± 0.018	Frequency of TM query	Query formulation
8.	0.384 ± 0.022	Frequency of TR query	Query formulation
9.	0.374 ± 0.028	Ratio searching to browsing	Search result interaction
10.	0.214 ± 0.027	Total query	Query formulation
11.	0.17 ± 0.029	Frequency of NL query	Query formulation
12.	0.157 ± 0.018	Ratio of relevant webpage from browsing to searching	Search result interaction
13.	0.164 ± 0.032	Total relevant webpage	Search result interaction
14.	0.127 ± 0.044	Total visited webpages	Search result interaction

No.	Average merit ± SD	Feature	Feature Set
15.	0.126 ± 0.018	Session length	Query formulation
16.	0.081 ± 0.01	First relevant result - query order type	Search result interaction
17.	0.041 ± 0.006	Average dwell time in visited webpage	Search result interaction
18.	0 ± 0	Average query length	Query formulation
19.	0 ± 0	Average visited webpage per query	Search result interaction

6.3 Model Development

The classification models on the dataset were developed to predict health topic familiarity of the participants. Weka [65] was utilized as the tool for providing the classifiers: BF Tree, Naïve Bayes, and Sequential Minimal Optimization (SMO). Standard settings were applied in BF Tree and Naïve Bayes classifiers. For SMO classifier, the kernel was set to Polynomial kernel and the complexity of parameter C was assigned to c = 2.0. The first classification process was conducted on three feature sets, i.e., query formulation feature set, search-result interaction feature set, and the combination of query formulation and search result interaction feature sets. The second classification process was conducted on top 12 features by information gain value. The top 12 feature sets from the highest to the lowest value were first query type, query reformulation pattern, search efficiency, average reading score of selected relevant web pages, successive query type, frequency of NM query, frequency of TM query, frequency of TR query, ratio of searching to browsing, total query, frequency of NL query, and ratio of relevant webpage from browsing to searching. Figure 6.2 shows the development of familiarity prediction model.

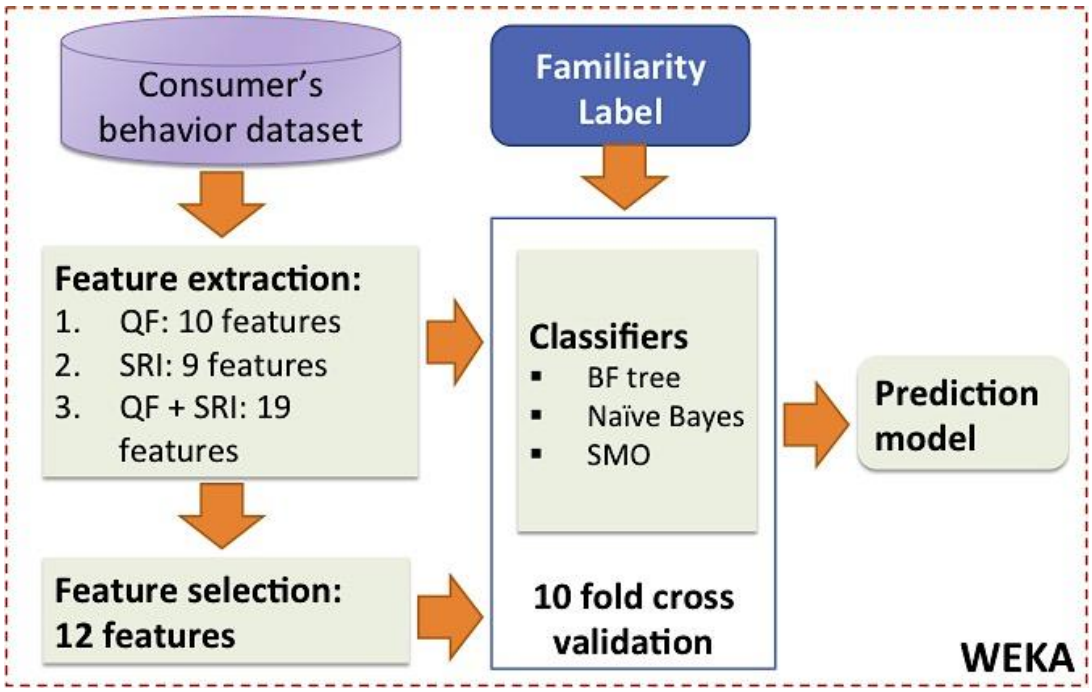


Figure 6.2 Model development

6.4 Result and Analysis

The performance of all classifiers on four different feature sets was compared in terms of accuracy at significance level of 0.05, as reported in Table 6.5. The features sets comprised of: query formulation set (10 features), search-result interaction set (9 features), the combination of query formulation set and search-result interaction set (19 features), and feature selection result set (12 features with the highest information gain score). All classifiers were evaluated in stratified ten-fold cross validation.

As shown in Table 6.5, query formulation feature set performs moderately well in predicting health topic familiarity. This finding indicates that the query formulation characterizes the searching behavior in different familiarity groups. Less familiar participants submitted more queries because they need to build their initial understanding with the searching topic before they can locate relevant information. One of the features in top 12 features selection result set is the first query type. This feature has the highest information gain value. Detail investigation on the first query type shows that the use of new health terminology

(both lay and medical) increases with the topic familiarity (5.7% for L1 participants, 17.67% for L2 participants, and 21.75% for L3 participants).

Table 6.5 Accuracy of the classifiers

Feature Sets	BF Tree (%)	Naïve Bayes (%)	SMO (%)
Query formulation (10 features)	73.33	80.00	81.67*
Search-result interaction (9 features)	70.83	68.33	72.50
Query formulation and search result interaction (19 features)	70.83	80.83	85.00*
Feature selection result (12 features)	73.33	87.50 *	90.83*

* indicates a statistically better performance than the baseline scheme (BF Tree)

The accuracy of all classifiers declined on the search result interaction feature set. There is no unique pattern between familiarity labels in several features (i.e., the total number of web pages visited, the number of relevant web pages, and the average number of visited web pages per query). In the features selection result, *search efficiency* and *average reading score of relevant web pages* have the highest scores of all features in the search result interaction set. More familiar participants achieved better search efficiency than the less familiar participants. Some of the familiar participants were able to discover relevant webpage from the first query submission and from the first selected retrieval result. On the average reading score of relevant web pages feature, familiar participants tend to choose more difficult web pages (e.g., a health article with 7 or more grade score and a medical paper from PubMed) than less familiar participants. The average reading scores were 6.34 (L1), 6.65 (L2), and 7.22 (L3).

Naïve Bayes and SMO classifiers achieved higher accuracy on the compounding of query formulation and search result interaction feature sets. It suggests that both feature sets are feasible to distinguish between unfamiliar, somewhat familiar, and familiar participants. Both classifiers achieved even higher accuracy on the feature selection result set. This result reaffirms the feasibility of query formulation and

search result interaction in characterizing health information search behavior. The advantage of obtaining higher accuracy with fewer features is the reduction of the dimensional problem in the real applications. Additional preprocessing to provide necessary personalization in information search should be done fast and efficiently.

This page is intentionally left blank.

Chapter 7

Summary

7.1 Conclusion

This thesis proposes individual health topic familiarity as a determinant factor in personalizing health information search. In health information search, the system is expected to present health information as closely as possible to the consumer's level of familiarity because the retrieved information influences health decision-making. The proposed health topic familiarity is validated by examining the effects of health topic familiarity on the health information search activity and developing the familiarity prediction model.

The result and findings from examining the effects of health topic familiarity on health information search confirm that health topic familiarity affects the participant's behavior in this study in terms of search activity pattern, query formulation, and search result interaction. The analysis of state transitions identified unique search pattern between different familiarity groups (unfamiliar, somewhat familiar, and familiar) in the designated demographic participants. The common patterns in unfamiliar group were frequent query modifications, with relatively low search efficiency, and accessing and evaluating selected results from a health website. The somewhat familiar group performed frequent query modifications, but with better search efficiency, and accessed and evaluated selected results from a health website. Finally, the familiar group successfully discovered relevant results from the first query submission, performed verification by accessing several health websites after they discovered relevant results, and directly accessed consumer health information websites.

Health information search systems can use the search pattern knowledge to analyze consumer's search behavior and to identify consumer's familiarity with the health topic. As the initial effort, this thesis presents the familiarity prediction

model based on the feature extraction from search activity pattern, i.e., query formulation and search result interaction. The performance result of the prediction model shows that the selected features are effective and feasible.

7.2 Implications for Health Information Search System Design

Results and findings form this thesis show that addressing individual familiarity in health information search systems is inevitable to improve the overall search process. The search system identifies the familiarity by analyzing the search activity pattern and provides the appropriate supports based on the consumer's familiarity.

To support unfamiliar consumers, health information search systems should implement assistive features during the construction of health queries and select understandable health information. These systems could help consumers to build queries using predefined diagnosis questionnaires and/or human anatomy diagrams. To support unfamiliar searchers with the identification of potentially relevant results, these systems should automatically extract a consumer-friendly definition of the submitted health query, adjust the rankings of the items retrieved, and suggest a related term using CHV. For more familiar searchers, these systems could be of assistance by locating additional relevant results. Based on the patterns exhibited in the present study, groups L2 and L3 were likely to continue the search process after they discovered relevant information. Systems could assist this process by clustering similar items into topic clusters in the page showing the results retrieved, by adjusting the ranking of retrieval items, and by providing a summary of health topic keywords.

7.3 Limitations and Future Studies

Most of the results obtained in this thesis correspond to thesis objectives, but a more comprehensive user study is required for further validation. The participants involved in the data collection shared several common demographic characteristics, i.e., higher education and a high level of experience in using the Internet. Therefore, the generalizability of the results is limited. A future user

study should extend the background of the participants. The next limitation is the time spent examining the results retrieved and evaluating the selected web pages were not considered in the search activities model. The time variable may characterize the search behaviors of different familiarity groups and it needs to be considered in future studies.

Another improvement area is in the familiarity prediction model. There are other factors characterizing health information search behavior that are not included in the features selection, such as the search route pattern, the time variables and the consumer's cognitive style. Those factors may have stronger effect on health topic familiarity and need to be explored in further study.

The findings of this thesis could contribute to the development of a more advanced personalized health information search system based on the individual's health topic familiarity. This type of system could identify the consumer's familiarity with health topics by analyzing their usage behavior to provide suitable support. Because health information search remains challenging for most consumers, this approach would be a major improvement in health information search systems.

Another contribution of this paper is the method of modeling search behavior as the sequence of search activities. This method can be applied to investigate the effects of other factors in health information search process. This method can also be utilized to model the search behavior in other domain with some modifications. Modification is required because the search activity types in this thesis are designed specifically for the search process in health domain.

This page is intentionally left blank.

References

[1] N. L. Atkinson, S. L. Saperstein, and J. Pleis, 'Using the Internet for Health-Related Activities: Findings From a National Probability Sample', *Journal of Medical Internet Research*, vol. 11, no. 1, Jan. 2009.

[2] F. Beck, J.-B. Richard, V. Nguyen-Thanh, I. Montagni, I. Parizot, and E. Renahy, 'Use of the Internet as a Health Information Resource Among French Young Adults: Results From a Nationally Representative Survey', *Journal of Medical Internet Research*, vol. 16, no. 5, Jan. 2014.

[3] A. Bianco, R. Zucco, C. G. A. Nobile, C. Pileggi, and M. Pavia, 'Parents Seeking Health-Related Information on the Internet: Cross-Sectional Study', *Journal of Medical Internet Research*, vol. 15, no. 9, Jan. 2013.

[4] S. Fox and M. Duggan, 'The social life of health information', *Pew Internet*, 15-Jan-2013. [Online]. Available: http://www.pewinternet.org/files/old-media//Files/Reports/PIP_HealthOnline.pdf. [Accessed: 27-Aug-2014].

[5] Y. Takahashi, T. Ohura, T. Ishizaki, S. Okamoto, K. Miki, M. Naito, R. Akamatsu, H. Sugimori, N. Yoshiike, K. Miyaki, T. Shimbo, and T. Nakayama, 'Internet Use for Health-Related Information via Personal Computers and Cell Phones in Japan: A Cross-Sectional Population-Based Survey', *Journal of Medical Internet Research*, vol. 13, no. 4, Jan. 2011.

[6] Q. T. Zeng and T. Tse, 'Exploring and Developing Consumer Health Vocabularies', *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 24–29, Jan. 2006.

[7] Q. T. Zeng, T. Tse, J. Crowell, G. Divita, L. Roth, and A. C. Browne, 'Identifying consumer-friendly display (CFD) names for health concepts', *AMIA Annual Symposium Proceedings*, pp. 859–863, 2005.

[8] G. Peterson, P. Aslani, and K. A. Williams, 'How do Consumers Search for and Appraise Information on Medicines on the Internet? A Qualitative Study Using Focus Groups', *Journal of Medical Internet Research*, vol. 5, no. 4, Jan. 2003.

[9] D. Soergel, T. Tony, and L. Slaughter, 'Helping healthcare consumers understand: an "'interpretive layer'"for finding and making sense of medical information', *Medinfo*, vol. 2, pp. 931–935, 2004.

[10] J. B. Walther, S. Pingree, R. P. Hawkins, and D. B. Buller, 'Attributes of Interactive Online Health Information Systems', *Journal of Medical Internet Research*, vol. 7, no. 3, Jan. 2005.

[11] Q. T. Zeng, J. Crowell, R. M. Plovnick, E. Kim, L. Ngo, and E. Dibble, 'Assisting Consumer Health Information Retrieval with Query Recommendations', *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 80–90, Jan. 2006.

[12] A. B. Can and N. Baykal, 'MedicoPort: A medical search engine for all',*Computer Methods and Programs in Biomedicine*, vol. 86, no. 1, pp. 73–86, Jan. 2007.

[13] G. Luo, C. Tang, H. Yang, and X. Wei, 'MedSearch', *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, Jan. 2008.

[14] Q. Zeng, E. Kim, J. Crowell, and T. Tse, 'A Text Corpora-Based Estimation of the Familiarity of Health Terminology', *Biological and Medical Data Analysis*, pp. 184–192, Jan. 2005.

[15] A. Keselman, T. Tse, J. Crowell, A. Browne, L. Ngo, and Q. Zeng, 'Assessing Consumer Health Vocabulary Familiarity: An Exploratory Study', *Journal of Medical Internet Research*, vol. 9, no. 1, Jan. 2007.

[16] Q. Zeng-Treitler, S. Goryachev, T. Tse, A. Keselman, and A. Boxwala, 'Estimating Consumer Familiarity with Health Terminology: A Context-based Approach', *Journal of the American Medical Informatics Association*, vol. 15, no. 3, pp. 349–356, Jan. 2008.

[17] G. Leroy, J. E. Endicott, D. Kauchak, O. Mouradi, and M. Just, 'User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention', *Journal of Medical Internet Research*, vol. 15, no. 7, Jan. 2013.

[18] S. Kandula, D. Curtis, and Q. Zeng-Treitler, 'A semantic and syntactic text simplification tool for health content', *AMIA Annual Symposium Proceedings*, pp. 366–370, Nov. 2010.

[19] M.-S. Paukkeri, M. Ollikainen, and T. Honkela, 'Assessing user-specific difficulty of documents', *Information Processing & Management*, vol. 49, no. 1, pp. 198–212, Jan. 2013.

[20] G. Leroy, E. Eryilmaz, and B. T. Laroya, 'Health information text characteristics', *AMIA Annual Symposium Proceedings*, vol. 2006, pp. 479–483, 2006.

[21] G. Luo, 'Design and Evaluation of the iMed Intelligent Medical Search Engine', *2009 IEEE 25th International Conference on Data Engineering*, Jan. 2009.

[22] P. Doupi and J. van der Lei, 'Towards personalized Internet health information: the STEPPS architecture', *Informatics for Health and Social Care*, vol. 27, no. 3, pp. 139–151, Jan. 2002.

[23] C. LeRouge, Jiao, S. Sneha, and K. Tolle, 'User profiles and personas in the design and development of consumer health technologies',*International Journal of Medical Informatics*, vol. 82, no. 11, Jan. 2013.

[24] S. Y. Rieh and H. Xie, 'Analysis of multiple query reformulations on the web: The interactive information retrieval context', *Information Processing & Management*, vol. 42, no. 3, pp. 751–768, Jan. 2006.

[25] B. J. Jansen, D. L. Booth, and A. Spink, 'Determining the informational, navigational, and transactional intent of Web queries', *Information Processing & Management*, vol. 44, no. 3, pp. 1251–1266, Jan. 2008.

[26] T. Saracevic, P. Kantor, A. Y. Chamis, and D. Trivison, 'A study of information seeking and retrieving: Background and methodology', in*Readings in Information Retrieval*, K. S. Jones and P. Willett, Eds. Morgan Kauffman Publishers, 1997.

[27] B. M. Wildemuth, 'The effects of domain knowledge on search tactic formulation', *Journal of the American Society for Information Science and Technology*, vol. 55, no. 3, pp. 246–258, Jan. 2004.

[28] B. J. Jansen, M. Zhang, and A. Spink, 'Patterns and transitions of query reformulation during web searching', *International Journal of Web Information Systems*, vol. 3, no. 4, pp. 328–340, Jan. 2007.

[29] R. Hu, K. Lu, and S. Joo, 'Effects of topic familiarity and search skills on query reformulation behavior', *Proceedings of the American Society for Information Science and Technology*, vol. 50, no. 1, pp. 1–9, Jan. 2013.

[30] R. W. White, S. T. Dumais, and J. Teevan, 'Characterizing the influence of domain expertise on web search behavior', *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, Jan. 2009.

[31] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag, 'Personalizing web search results by reading level',*Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, Jan. 2011.

[32] S. Bhavnani, 'Important cognitive components of domain-specific search knowledge', *Ann Arbor 1001*, vol. 2001, pp. 48109–1092, 2001.

[33] A. Aula, N. Jhaveri, and M. Käki, 'Information search and re-access strategies of experienced web users', *Proceedings of the 14th international conference on World Wide Web - WWW '05*, Jan. 2005.

[34] A. Thatcher, 'Web search strategies: The influence of Web experience and task type', *Information Processing & Management*, vol. 44, no. 3, pp. 1308–1329, Jan. 2008.

[35] I. Xie and S. Joo, 'Transitions in search tactics during the Web-based search process', *Journal of the American Society for Information Science and Technology*, vol. 61, no. 11, pp. 2188–2205, Jan. 2010.

[36] P. Boldi, F. Bonchi, C. Castillo, and S. Vigna, 'From "Dango" to "Japanese Cakes": Query Reformulation Models and Patterns', *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Jan. 2009.

[37] H. A. Hembrooke, L. A. Granka, G. K. Gay, and E. D. Liddy, 'The effects of expertise and feedback on search term selection and subsequent learning', *Journal of the American Society for Information Science and Technology*, vol. 56, no. 8, pp. 861–871, Jan. 2005.

[38] K. Kinley, D. Tjondronegoro, H. Partridge, and S. Edwards, 'Modeling users' web search behavior and their cognitive styles', *Journal of the Association for Information Science and Technology*, vol. 65, no. 6, pp. 1107–1123, Jan. 2014.

[39] D. Kelly and C. Cool, 'The effects of topic familiarity on information search behavior', *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries - JCDL '02*, Jan. 2002.

[40] M. J. Cole, J. Gwizdka, C. Liu, R. Bierig, N. J. Belkin, and X. Zhang, 'Task and user effects on reading patterns in information search',*Interacting with Computers*, vol. 23, no. 4, pp. 346–362, Jan. 2011.

[41] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais, 'Characterizing web content, user interests, and search behavior by reading level and topic', *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, Jan. 2012.

[42] I. Xie, *Interactive information retrieval in digital environments*. United States: IGI Publishing, 2008.

[43] A. A. Shiri and C. Revie, 'The effects of topic complexity and familiarity on cognitive and physical moves in a thesaurus-enhanced search environment', *Journal of Information Science*, vol. 29, no. 6, pp. 517–526, Jan. 2003.

[44] J. Lin and J. W. Wilbur, 'Modeling actions of PubMed users with n-gram language models', *Information Retrieval*, vol. 12, no. 4, pp. 487–503, Jan. 2008.

[45] P. Vakkari, M. Pennanen, and S. Serola, 'Changes of search terms and tactics while writing a research proposal', *Information Processing & Management*, vol. 39, no. 3, pp. 445–463, Jan. 2003.

[46] I. Xie and S. Joo, 'Factors affecting the selection of search tactics: Tasks, knowledge, process, and systems', *Information Processing & Management*, vol. 48, no. 2, pp. 254–270, Jan. 2012.

[47] B. Allen, 'Topic Knowledge and Online Catalog Search Formulation',*The Library Quarterly*, vol. 61, no. 2, pp. 188–213, Jan. 1991.

[48] Z. Yue, S. Han, and D. He, 'A Comparison of Action Transitions in Individual and Collaborative Exploratory Web Search', *Lecture Notes in Computer Science*, pp. 52–63, Jan. 2012.

[49] H.-M. Chen and M. D. Cooper, 'Stochastic modeling of usage patterns in a web-based information system', *Journal of the American Society for Information Science and Technology*, vol. 53, no. 7, pp. 536–548, Jan. 2002.

[50] 'Home - European Patients Academy on Therapeutic Innovation'. [Online]. Available: http://www.patientsacademy.eu/index.php/en/. [Accessed: 20-Nov-2014].

[51] ‘The determinants of health’, *World Health Organization*, 01-Dec-2010. [Online]. Available: http://www.who.int/hia/evidence/doh/en/. [Accessed: 15-Dec-2014].

[52] R. W. White and E. Horvitz, ‘Cyberchondria’, *ACM Transactions on Information Systems*, vol. 27, no. 4, pp. 1–37, Jan. 2009.

[53] A. Keselman, L. Massengale, L. Ngo, A. Browne, and Q. Zeng, ‘The Effect of User Factors on Consumer Familiarity with Health Terms: Using Gender as a Proxy for Background Knowledge About Gender-Specific Illnesses’, *Biological and Medical Data Analysis*, pp. 472–481, Jan. 2006.

[54] G. H. McLaughlin, ‘SMOG grading: A new readability formula’, *Journal of reading*, vol. 12, 8 vols., pp. 639–646, 1969.

[55] P. Fitzsimmons, B. Michael, J. Hulley, and G. Scott, ‘A readability assessment of online Parkinson’s disease information’, *The Journal of the Royal College of Physicians of Edinburgh*, vol. 40, no. 4, pp. 292–296, Jan. 2010.

[56] ‘Health Topics: Common Health Topics’, *Yahoo Health*. [Online]. Available: http://health.yahoo.net/directory/health-channels. [Accessed: 21-Jan-2014].

[57] R. M. Parker, D. W. Baker, M. V. Williams, and J. R. Nurss, ‘The test of functional health literacy in adults’, *Journal of General Internal Medicine*, vol. 10, no. 10, pp. 537–541, Jan. 1995.

[58] ‘Free Screen Recording Software’, *Camstudio*. [Online]. Available: http://camstudio.org/. [Accessed: 08-Oct-2013].

[59] A. Stolcke, ‘SRILM-an extensible language modeling toolkit’,*Proceedings of the 7th International Conference on Spoken Language Processing*, 2002.

[60] I. H. Witten and T. C. Bell, ‘The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression’, *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1085–1094, Jan. 1991.

[61] S. F. Chen, D. Beeferman, and R. Rosenfield, ‘Evaluation metrics for language models’, *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 275–280, 1998.

[62] J. Liu, C. Liu, and N. Belkin, ‘Examining the effects of task topic familiarity on searchers’ behaviors in different task types’, *Proceedings of the American Society for Information Science and Technology*, vol. 50, no. 1, pp. 1–10, Jan. 2013.

[63] X. Zhang, H. G. B. Anghelescu, and X. Yuan, 'Domain knowledge, search behavior, and search effectiveness of engineering and science students: An exploratory study', *Information research*, vol. 10, 2 vols., 2005.

[64] I. Puspitasari, K. Moriyama, K. Fukui, and M. Numao, 'Effects of Individual Health Topic Familiarity on the Activity Pattern during Health Information Searches', *JMIR Med Inform (forthcoming)*, 2015;3(1):e16. doi: 10.2196/medinform.3803. PMID: 25783222.

[65] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutmann, and I. H. Witten, 'The WEKA Data Mining Software: An Update', *SIGKDD Explorations*, vol. 11, 1 vols., 2009.

[66] M. A. Hall and G. Holmes, 'Benchmarking attribute selection techniques for discrete class data mining', *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 1437–1447, Jan. 2003.

This page is intentionally left blank.

Appendix A
Health Terminology Familiarity Questionnaire

Section 1

Instruction

- For each item below, check one option that is **most closely related** to the *italic* word.
- If you had never heard the *italic* word/phrase before, please **do not guess** the answer and **select the unknown** option.

1. *Eczema*

☐ Skin Inflammation ☐ Broken bone ☐ Unknown
☐ Stomach problem ☐ Movement Disorder

2. *Topical ointment*

☐ Protein ☐ Surgery ☐ Unknown
☐ Hormone ☐ Medicine

3. *Automated External Defibrillator*

☐ First aid medical device ☐ Body scanner device
☐ Medical measurement device ☐ Unknown
☐ Surgical instrument

4. *Heart attack*

☐ Heart and blood vessel disease ☐ Endocrine disease
☐ Digestive system disease ☐ Unknown
☐ Kidney disease

5. *Urinalysis*

☐ Disease ☐ Hormone ☐ Unknown
☐ Medical test ☐ Enzyme

6. *Urine specific gravity*

☐ pH test ☐ Protein test ☐ Unknown
☐ Concentration test ☐ Glucose test

7. *Cholesterol*

☐ Food allergy ☐ Disease ☐ Unknown
☐ Fat substance ☐ Protein

8. *Simvastatin*

☐ Surgical procedure ☐ Vaccine ☐ Unknown
☐ Detoxification ☐ Oral medication

Section 2

Instruction: Select one correct answer for each question below.

If you had never heard the *italic* word/phrase before, please **do not guess** the answer and **select the unknown** option.

1. If you are diagnosed with *eczema*, it means that …
   - ☐ your skin shows patches of itchy, redness, and thickened area
   - ☐ you have a broken bone in your wrist
   - ☐ the inside lining of your stomach is wounded
   - ☐ your tendon or muscle in the knee joint is injured
   - ☐ Unknown

2. A *topical ointment* is …
   - ☐ a type of protein for building muscle tissue and repairing damaged tissues
   - ☐ a body chemical that responses to allergy or infection
   - ☐ a diagnostic test involving the removal of sample tissue
   - ☐ a type medication that is applied to the skin to reduce the inflammation
   - ☐ Unknown

3. The medical kit box (AED) as in the picture below is located in many public places. This kit is used to …

   - ☐ deliver electric shocks to a patient's heart in a sudden heartbeat stop incident
   - ☐ measure blood sugar level
   - ☐ filter harmful waste, salt, and excess fluid from the blood
   - ☐ determine the severity of injuries by scanning the affected parts
   - ☐ Unknown

4. You frequently find brochures in the clinic or hospital about how to recognize a *heart attack*. To you heart attack means …
   - ☐ the heart suddenly stops beating unexpectedly
   - ☐ the artery that carries blood to the heart is blocked
   - ☐ heartbeat rhythm problem, the heart beats too fast, too slow, or too irregularly
   - ☐ a damage to the heart muscle
   - ☐ Unknown

5. *Urinalysis* refers to …
   - ☐ disorder in the kidney and urine tract system
   - ☐ medical test that examines the physical, chemical, and microscopic properties of urine
   - ☐ hormone system that regulates the balance of blood pressure and water
   - ☐ enzyme that breaks down protein into smaller particle, e.g. amino acid
   - ☐ Unknown

6. *Urine density test* measures …
   - ☐ the level of acid in a urine
   - ☐ the concentration of substances in a urine
   - ☐ the excess amount of protein found in a urine sample
   - ☐ the amount of sugar found in a urine sample
   - ☐ Unknown

7. Too much *bad cholesterol* in the blood is dangerous because …
   - ☐ it may damage liver
   - ☐ it leads to kidney stone formation within the kidney or in the urinary tract
   - ☐ it can obstruct the absorption of good nutrients in the small intestine
   - ☐ it leads to artery blockage and increases heart attack risk
   - ☐ Unknown

8. *Simvastatin* is mainly prescribed for …
   - ☐ reducing total cholesterol level
   - ☐ the treatment of mild to moderate pain, inflammation and fever
   - ☐ lowering blood pressure level
   - ☐ the treatment of nasal congestion and runny nose from allergy
   - ☐ Unknown

Section 3

Instruction: Select one correct answer for each question below.

If you had never heard the *italic* word/phrase before, please **do not guess** the answer and **select the unknown** option.

1. If you are diagnosed with *atopic dermatitis*, it means that …
   - ☐ the inside lining of your stomach is wounded
   - ☐ your skin shows patches of itchy, redness, and thickened area
   - ☐ your tendon or tissue in the knee joint is injured
   - ☐ you have fracture(s) in your wrist
   - ☐ Unknown
2. A *topical corticosteroid* is …
   - ☐ a type of protein for building muscle tissue and repairing damaged tissues
   - ☐ a body chemical that responses to allergy or infection
   - ☐ a diagnostic test that involves taking a sample of tissue for an examination under a microscope
   - ☐ a type of drug to reduce inflammation and thickening of the skin
   - ☐ Unknown
3. An Automated External Defibrillator is a portable device to …
   - ☐ deliver electric shocks to a patient's heart in a sudden cardiac arrest incident
   - ☐ measure the approximate concentration of glucose in the blood
   - ☐ filter harmful waste, salt, and excess fluid from the blood
   - ☐ determine the severity of injuries by scanning the affected body parts
   - ☐ Unknown
4. You frequently find brochures in the clinic or hospital about how to recognize a *myocardial infarction (MI)*. MI means …
   - ☐ a heart condition in which the heart suddenly and unexpectedly stops beating
   - ☐ a blockage in the artery that carries blood to the heart
   - ☐ heartbeat rhythm problem, the heart may beat too fast, too slow, or too irregularly
   - ☐ a damage to the heart muscle
   - ☐ Unknown

5. *Routine and Microscopy (R&M)* refers to …
   - ☐ disorder in the kidney and urine tract system
   - ☐ medical test that examines the physical, chemical, and microscopic properties of urine
   - ☐ hormone system that regulates the balance of blood pressure and water
   - ☐ enzyme that breaks down protein into smaller particle, e.g. amino acid
   - ☐ Unknown

6. *Urine specific gravity measures* …
   - ☐ how acidic or alkaline the urine is
   - ☐ the concentration of all chemical particles in the urine
   - ☐ the excess amount of protein found in a urine sample
   - ☐ the amount of glucose found in a urine sample
   - ☐ Unknown

7. High level of *low-density lipoprotein* may cause …
   - ☐ damage to the liver
   - ☐ the formation of kidney stone within the kidney or in the urinary tract
   - ☐ disorder in small intestine function to absorb good nutrients from food
   - ☐ artery blockage that can increase coronary disease risk
   - ☐ Unknown

8. *Statins* or *HMG COa Reductase Inhibitors* are drugs used to …
   - ☐ reduce blood cholesterol level
   - ☐ treat mild to moderate pain, inflammation and fever
   - ☐ lower blood pressure level
   - ☐ relieve nasal congestion and reduce the symptoms of an allergic reaction
   - ☐ Unknown

Appendix B
Transition between Search Activity Types

Transition probability matrix in Group L1 (0.0% - 100.0%)

From	To	AccSE	AccHW	NewQ	ModQ	ExamSR	DisSR	SelHI	SelGI	EvalI	AccF	AccB	FindQ	UseHI	UseGI	DisHI	DisGI	UnchkHI	UnchkGI
AccSE			38.3	61.7															
AccHW			25.0	75.0															
NewQ		1.3			96.2	2.6													
ModQ					97.5	0.6			1.2	0.6									
ExamSR			0.3	1.3		10.4		63.1	22.8		0.3	1.7							
DisSR		20.6		2.9	67.6					2.9	2.9	2.9							
SelHI		0.7			2.7	0.7		23.7	9.5	61.4		0.3			0.3			0.3	0.3
SelGI					2.8	0.9		32.7	7.5	54.2									1.9
EvalI		2.7	0.2			1.1				8.0	15.6	5.3	2.7	21.2	6.2	28.3	8.8		
AccF		2.3								70.5	18.6	3.9	1.6	0.8		1.6		0.8	
AccB										88.9	6.9	2.8				1.4			
FindQ										15.0	10.0		15.0	30.0		20.0	10.0		
UseHI		24.8	1.8			8.3		0.9		21.1	5.5	10.1	0.0	8.3	1.8	7.3	0.9	6.4	2.8
UseGI		48.5				12.1				12.1	3.0		3.0	6.1		3.0		6.1	6.1
DisHI		22.8	0.6			18.9		0.6	0.6	37.8	0.6	7.2				5.0	1.1	3.9	1.1
DisGI		27.3				14.5				41.8		3.6				3.6	7.3	1.8	
UnchkHI		28.6				4.8				14.3		9.5			4.8			28.6	9.5
UnchkGI		40.0																60.0	

Transition probability matrix in Group L2 (0.0% - 100.0%)

From	To: AccSE	AccHW	NewQ	ModQ	ExamSR	DisSR	SelHI	SelGI	EvalI	AccF	AccB	FindQ	UseHI	UseGI	DisHI	DisGI	UnchkHI	UnchkGI
AccSE			54.8	45.2														
AccHW				100.0														
NewQ	1.5				98.5													
ModQ					100.0													
ExamSR			1.2	1.2		11.2	66.5	18.8			1.2							
DisSR				88.2					5.9		5.9							
SelHI				1.3			25.2	9.4	63.5				0.6					
SelGI				1.9	1.9		11.5	9.6	75.0									
EvalI	2.7				0.8				10.0	10.0	6.5	5.0	36.2	5.0	13.8	9.6		0.4
AccF	2.9								88.2	5.9							2.9	
AccB									96.4				3.6					
FindQ	6.7											13.3	33.3	13.3	33.3			
UseHI	38.0				6.3				22.8	6.3	6.3		3.8	1.3	2.5	1.3	5.1	6.3
UseGI	21.4			7.1	14.3				21.4				14.3		14.3		7.1	
DisHI	30.8	2.6			28.2				30.8		2.6				5.1			
DisGI	33.3				**37.5**				**12.5**		**8.3**		**4.2**				**4.2**	
UnchkHI	33.3	33.3			33.3													
UnchkGI	**50.0**												**50.0**					

Transition probability matrix in Group L3 (0.0% - 100.0%)

From	To: AccSE	AccHW	NewQ	ModQ	ExamSR	DisSR	SelHI	SelGI	EvalI	AccF	AccB	FindQ	UseHI	UseGI	DisHI	DisGI	UnchkHI	UnchkGI
AccSE			57.8	42.2														
AccHW			75.0	25.0														
NewQ	1.8				98.2													
ModQ					91.7	2.1	6.3											
ExamSR						6.9	77.9	12.2	0.8		2.3							
DisSR	10.0			70.0	10.0				10.0									
SelHI				1.9			27.8	5.7	63.9				0.6					
SelGI							18.5	7.4	74.1									
EvalI	1.9				1.0		0.5		8.7	10.1	1.9	3.4	38.9	3.4	24.5	5.8		
AccF	3.6								85.7	10.7								
AccB							20.0		80.0									
FindQ												22.2	55.6		22.2			
UseHI	23.1	1.5			21.5				26.2	3.1			7.7	1.5	6.2	3.1	3.1	3.1
UseGI									33.3				33.3			33.3		
DisHI	28.3	5.7			20.8		1.9		26.4	3.8	5.7		3.8				1.9	1.9
DisGI	35.7				21.4				14.3				7.1			14.3	7.1	
UnchkHI	100.0																	
UnchkGI									50.0								50.0	

This page is intentionally left blank.

Appendix C
Most Frequent 5-gram Sequence Patterns in Each Familiarity Group

Top 20 frequent 5-gram sequence patterns in group L1

Rank.	Pattern	Occurrence Frequency	Percentage (%)[a]
1	A:AccSE A:ModQ E:ExamSR A:SelHI E:EvalI	55	5.8
2	E:EvalI D:DisHI A:AccSE A:ModQ E:ExamSR	35	3.7
3	A:ModQ E:ExamSR A:SelHI E:EvalI U:UseHI	27	2.9
4	E:EvalI D:DisHI E:ExamSR A:SelHI E:EvalI	26	2.8
5	A:ModQ E:ExamSR A:SelHI E:EvalI D:DisHI	24	2.6
6	D:DisHI A:AccSE A:ModQ E:ExamSR A:SelHI	23	2.4
7	E:EvalI U:UseHI A:AccSE A:ModQ E:ExamSR	23	2.4
8	E:EvalI A:XplorF E:EvalI A:XplorF E:EvalI	22	2.3
9	E:EvalI D:DisHI E:EvalI D:DisHI E:EvalI	22	2.3
10	E:ExamSR A:SelHI E:EvalI D:DisHI A:AccSE	22	2.3
11	E:ExamSR A:SelHI E:EvalI D:DisHI E:ExamSR	21	2.2
12	A:SelHI E:EvalI D:DisHI A:AccSE A:ModQ	20	2.1
13	A:SelHI E:EvalI D:DisHI E:ExamSR A:SelHI	18	1.9
14	A:AccSE A:NewQ E:ExamSR A:SelHI A:SelHI	17	1.8
15	E:ExamSR A:SelHI E:EvalI U:UseHI A:AccSE	17	1.8
16	A:SelHI E:EvalI U:UseHI A:AccSE A:ModQ	17	1.8
17	D:DisHI E:ExamSR A:SelHI E:EvalI D:DisHI	16	1.7
18	U:UseHI A:AccSE A:ModQ E:ExamSR A:SelHI	16	1.7
19	A:AccSE A:NewQ E:ExamSR A:SelHI E:EvalI	15	1.6
20	A:ModQ E:ExamSR E:DisSR A:ModQ E:ExamSR	15	1.6
Total Top 20 frequent patterns		451	48.0

[a] Total number of all 5-gram sequences = 940

Top 20 frequent 5-gram sequence patterns in group L2

Rank.	Pattern	Occurrence Frequency	Percentage (%)[a]
1	A:AccSE A:ModQ E:ExamSR A:SelHI E:EvalI	24	5.4
2	E:EvalI U:UseHI A:AccSE A:ModQ E:ExamSR	24	5.4
3	A:ModQ E:ExamSR A:SelHI E:EvalI U:UseHI	23	5.2
4	A:AccSE A:NewQ E:ExamSR A:SelHI E:EvalI	21	4.7
5	U:UseHI A:AccSE A:ModQ E:ExamSR A:SelHI	21	4.7
6	E:ExamSR A:SelHI E:EvalI U:UseHI A:AccSE	20	4.5
7	A:SelHI E:EvalI U:UseHI A:AccSE A:ModQ	17	3.8
8	A:NewQ E:ExamSR A:SelHI E:EvalI U:UseHI	16	3.6
9	A:AccSE A:NewQ E:ExamSR A:SelHI A:SelHI	13	2.9
10	A:AccSE A:ModQ E:ExamSR A:SelHI A:SelHI	10	2.3
11	A:AccSE A:NewQ E:ExamSR A:SelGI E:EvalI	10	2.3
12	E:ExamSR E:DisSR A:ModQ E:ExamSR A:SelHI	10	2.3
13	E:ExamSR A:SelHI A:SelHI E:EvalI E:EvalI	10	2.3
14	A:ModQ E:ExamSR E:DisSR A:ModQ E:ExamSR	9	2.0
15	E:ExamSR A:SelHI A:SelHI A:SelHI E:EvalI	8	1.8
16	A:ModQ E:ExamSR A:SelGI E:EvalI D:DisGI	8	1.8
17	A:ModQ E:ExamSR A:SelHI E:EvalI D:DisHI	8	1.8
18	A:NewQ E:ExamSR A:SelHI A:SelHI E:EvalI	8	1.8
19	A:AccSE A:ModQ E:ExamSR E:DisSR A:ModQ	7	1.6
20	A:AccSE A:ModQ E:ExamSR A:SelGI E:EvalI	7	1.6
Total Top 20 frequent patterns		274	61.8

[a] Total number of all 5-gram sequences = 444

Top 20 frequent 5-gram sequence patterns in group L3

Rank.	Pattern	Occurrence Frequency	Percentage (%)[a]
1	A:AccSE A:NewQ E:ExamSR A:SelHI E:EvalI	18	5.0
2	A:AccSE A:ModQ E:ExamSR A:SelHI E:EvalI	17	4.7
3	A:NewQ E:ExamSR A:SelHI E:EvalI U:UseHI	16	4.5
4	E:EvalI D:DisHI A:AccSE A:ModQ E:ExamSR	12	3.3
5	A:ModQ E:ExamSR A:SelHI E:EvalI U:UseHI	11	3.1
6	A:AccSE A:NewQ E:ExamSR A:SelHI A:SelHI	10	2.8
7	E:ExamSR A:SelHI E:EvalI U:UseHI E:ExamSR	10	2.8
8	E:ExamSR A:SelHI A:SelHI A:SelHI E:EvalI	10	2.8
9	U:UseHI A:AccSE A:ModQ E:ExamSR A:SelHI	10	2.8
10	D:DisHI A:AccSE A:ModQ E:ExamSR A:SelHI	9	2.5
11	E:EvalI U:UseHI A:AccSE A:ModQ E:ExamSR	9	2.5
12	E:ExamSR A:SelHI E:EvalI D:DisHI A:AccSE	9	2.5
13	A:SelHI E:EvalI D:DisHI A:AccSE A:ModQ	9	2.5
14	A:SelHI E:EvalI U:UseHI A:AccSE A:ModQ	9	2.5
15	A:SelHI E:EvalI U:UseHI E:ExamSR A:SelHI	9	2.5
16	A:AccHW A:NewQ E:ExamSR A:SelHI E:EvalI	8	2.2
17	E:ExamSR A:SelHI E:EvalI A:XplorF E:EvalI	8	2.2
18	E:ExamSR A:SelHI E:EvalI U:UseHI A:AccSE	8	2.2
19	A:AccSE A:ModQ E:ExamSR A:SelHI A:SelHI	7	2.0
20	A:SelHI A:SelHI E:EvalI D:DisHI E:EvalI	7	2.0
Total Top 20 frequent patterns		206	57.4

[a] Total number of all 5-gram sequences = 359