

Knowledge Driven Search Intent Mining

A thesis submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

By

ASHUTOSH JADHAV
B.E., Veermata Jijabai Technological Institute, 2006
M.S., Wright State University, 2008

2016
Wright State University
Dayton, Ohio 45435

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

April 18, 2016

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Ashutosh Jadhav ENTITLED Knowledge Driven Search Intent Mining BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Amit Sheth, Ph. D.
Dissertation Director

Michael Raymer, Ph.D.
Director, Computer Science and Engineering
Ph.D. Program

Robert E.W. Fyffe, Ph.D.
Vice President for Research and Dean of the
Graduate School

Committee on
Final Examination

Professor Amit Sheth, Ph.D.

Professor Krishnaprasad Thirunarayan, Ph.D.

Professor Michael Raymer, Ph.D.

Professor Jyotishman Pathak, Ph.D.

ABSTRACT

Jadhav, Ashutosh. Ph.D. Department of Computer Science and Engineering, Wright State University, 2016. Knowledge Driven Search Intent Mining.

Rich background knowledge from biomedical knowledge bases and Wikipedia enables development of effective methods for: I) Intent mining from health-related search queries in disease agnostic manner II)Efficient browsing of informative health information shared on social media.

Understanding users' latent intents behind search queries is essential for satisfying a user's search needs. Search intent mining can help search engines to enhance its ranking of search results, enabling new search features like instant answers, personalization, search result diversification, and the recommendation of more relevant ads. Hence, there has been increasing attention on studying how to effectively mine search intents by analyzing search engine query logs. While state-of-the-art techniques can identify the domain of the queries (e.g. sports, movies, health), identifying domain-specific intent is still an open problem. Among all the topics available on the Internet, health is one of the most important in terms of impact on the user and forms one of the most frequently searched areas. This dissertation presents a knowledge-driven approach for domain-specific search intent mining with a focus on health-related search queries.

First, we identified 14 consumer-oriented health search intent classes based on inputs from focus group studies and based on analyses of popular health websites, literature surveys, and an empirical study of search queries. We defined the problem of classifying millions of health search queries into zero or more intent classes as a multi-label classification problem. Popular machine learning approaches for multi-label classification tasks (namely, problem transformation and algorithm adaptation methods) were not feasible due to the limitation of label data creations and health domain constraints. Another challenge in solving the search intent identification problem was mapping terms used by laymen to medical terms. To address these challenges, we developed a semantics-driven,

rule-based search intent mining approach leveraging rich background knowledge encoded in Unified Medical Language System (UMLS) and a crowd sourced encyclopedia (Wikipedia). The approach can identify search intent in a disease-agnostic manner and has been evaluated on three major diseases.

While users often turn to search engines to learn about health conditions, a surprising amount of health information is also shared and consumed via social media, such as public social platforms like Twitter. Although Twitter is an excellent information source, the identification of informative tweets from the deluge of tweets is the major challenge. We used a hybrid approach consisting of supervised machine learning, rule-based classifiers, and biomedical domain knowledge to facilitate the retrieval of relevant and reliable health information shared on Twitter in real time. Furthermore, we extended our search intent mining algorithm to classify health-related tweets into health categories. Finally, we performed a large-scale study to compare health search intents and features that contribute in the expression of search intent from more than 100 million search queries from smart devices (smartphones or tablets) and personal computers (desktops or laptops).

Contents

1	Introduction	1
1.1	Selection of Domain	4
1.2	Knowledge-driven Health Search Intent Mining	5
1.3	Twitter, a Health Information Source	11
1.4	Comparative Analysis of Effects of Device on Expression of Search Intents	14
1.5	Dissertation Organization	15
2	Search Intent Mining	16
2.1	Background	17
2.2	Search Intent Mining based on Query Log	17
2.3	Search Intent Mining for Personalization	18
2.4	Search Intent Mining Based on Click-through Data	19
2.5	Search Intent Mining for Query Classification	20
2.6	Search Intent Mining for Vertical Selection	21
3	Domain Specific Search Intent Mining	23
3.1	Motivation	23
3.1.1	Real-world Challenges with Mayo Clinic’s Consumer Health Information Portal	23
3.1.2	Opportunities and Challenges in Health Domain	25
3.2	Health Search Intent	27

3.2	Health Search Intent	27
3.3	Selection of Health Search Intent Classes	28
3.3.1	Online Health Information Searching: A Qualitative Approach for Exploring Consumer Perspectives	28
3.3.2	Analysis of Health Categories on Popular Websites	30
3.3.3	Survey of Health Information Seeking Literature	31
3.3.4	Empirical Study of Health Queries	32
3.4	Problem Statement	35
3.5	Multi-label Classification	35
3.5.1	Problem Transformation Methods	36
3.5.2	Algorithm Adaptation Methods	38
3.5.3	Challenges and Limitations	39
3.6	Knowledge-driven Approach	40
3.6.1	Biomedical Knowledge Bases	42
3.6.2	In the Context of Health Search Intent Mining	45
3.7	Concept Identification	45
3.7.1	Medical Concept Identification Tools	47
3.7.2	Concept Identification using MetaMap	48
3.7.3	Concept Identification Challenge	50
3.8	Consumer Health Vocabulary Generation Using Wikipedia	55
3.8.1	Approach	58
3.8.1.1	Candidate Pair Generation from Health Related Wikipedia Articles	59
3.8.1.2	Identification of CHV and medical terms from candidate pairs	62
3.9	The Corpus	64
3.9.1	Rationale for Data Selection	64
3.9.2	Data Source	65

3.9.3	Dataset Creation	65
3.9.4	Gold Standard Dataset Creation	66
3.10	Data Preprocessing	66
3.11	Classification Approach	69
3.11.1	Classification Rules	69
3.11.2	Classification Algorithm	72
3.12	Evaluations and Results	72
3.12.1	Classification Approach Evaluation	72
3.12.2	Classification Evaluation by Intent Classes	74
3.12.3	Distribution of Search Queries by Number of Classified Intent Classes	76
3.12.4	Evaluation with respect to three chronic diseases	76
3.13	Conclusion	79
4	A Hybrid Approach for Identification of Informative Tweets and Social Health Signals System	80
4.1	Introduction	81
4.2	Approach	84
4.2.1	Data Collection	84
4.2.2	Rule-based Filtering	84
4.2.3	Classification	86
4.2.4	Classification Features	88
4.3	Experiments and Evaluations	89
4.4	Social Health Signals System	91
4.4.1	Data Processing Pipeline	92
4.4.2	Question and Answering on Twitter data	93
4.4.3	Semantic Categorization	94

4.4.4	Social Health Signals User Interface	94
5	Evaluating the Process of Online Health Information Searching: A Qualitative Approach to Exploring Consumer Perspectives	96
5.1	Introduction	96
5.2	Methods	99
5.2.1	Study Participants and Recruitment	99
5.2.2	Data Collection and Analysis	100
5.3	Results	100
5.3.1	Overview	100
5.3.2	Motivations for Online Health Searching	100
5.3.3	Searching Strategies and Techniques	104
5.3.4	Content Preferences	106
5.4	Discussion	108
5.4.1	Principal Findings	108
5.4.2	Limitations	109
5.5	Conclusion	110
6	Comparative Analysis of Expressions of Search Intents From Personal Computers and Smart Devices	112
6.1	Introduction	112
6.1.1	Significance of Current Study	114
6.2	Methods	115
6.2.1	Data Source	115
6.2.2	Dataset Creation	116
6.2.3	Data Analysis	117
6.3	Results	121

6.4	Discussion	129
6.4.1	Overview	129
6.4.2	Principal Results	130
6.4.3	Comparison With Related Work	132
6.4.4	Limitations	132
6.4.5	Future Work	133
6.5	Conclusion	133
7	Conclusions	135
7.1	Summary	135
7.2	Future Directions	138
	Bibliography	143

List of Figures

1.1	Structured information provided by Google search for a ‘type 2 diabetes’ query	3
1.2	a snippet from a Wikipedia article on “knee effusion”.	10
3.1	MetaMap concept mapping for “stomach pain”. The MetaMap maps “stomach pain” to the concept “stomach ache” and the Semantic Type “Sign or Symptom”.	49
3.2	MetaMap concept mapping for “water in brain”.	50
3.3	MetaMap correct concept mapping for “water in brain”.	53
3.4	MetaMap correct concept mapping for “water on the knee”.	53
3.5	a snippet from a Wikipedia article on “knee effusion”.	54
3.6	Wikipedia category hierarchy.	57
3.7	Parent categories on Wikipedia category hierarchy	58
3.8	Approach for generating CHV	60
3.9	Wikipedia formatting patterns that are used to extracts candidate pairs	62
3.10	Search query data collection at Mayo Clinic	65
3.11	Hadoop-MapReduce framework with 16 nodes for MetaMap implementation	68
3.12	Functional overview of a Mapper	68
3.13	Classification rule for Drugs and Medications intent class	71
4.1	Social Health Signals Architecture	92
4.2	Social Health Signals User Interface	95

6.1	Screenshot of Mayo Clinic website for Diabetes (left-side box highlights organization of health information based on health categories	118
6.2	Distribution of the search queries by health categories.	122
6.3	Distribution of the search queries by number of words and number	124
6.4	Distribution of the search queries by number of characters.	124
6.5	Types of health search queries (how health information need is expressed).	126
6.6	Distribution of the search queries based on type of wh-questions.	126
6.7	Distribution of the search queries based on type of yes/no questions.	127
7.1	Medical question posted by a layman on one medical question-answering forum (DailyStrength)	139
7.2	Information extraction using search intent mining algorithm	140
7.3	Structured medical information extracted from unstructured medical question using techniques used in search intent mining algorithm.	140

List of Tables

3.1	List of health categories on popular health websites	32
3.2	List of health intent classes and their description with examples	34
3.3	Candidate term pairs from Wikipedia snippets	59
3.4	Evaluation of the classification approach	74
3.5	Performance of the classification approach with respect to individual intent classes .	75
3.6	Classification of search queries by intent classes	77
3.7	Classification of search queries by intent classes	78
3.8	Performance of the classification approach with respect to three major chronic diseases	78
4.1	Rule-based filtering	87
4.2	Performance of different classifiers in the informative tweet classification task	89
4.3	Classification performance with different combinations of the features	90
5.1	Characteristics of patients (n=19).	101
6.1	Categorization of health search queries based on the information mentioned in the queries such as gender, age group, and temporal information (June 2011-May 2013) .	122
6.2	Usage of query operators and special characters (June 11-May 13).	125
6.3	Linguistic analysis of health search queries (June 2011-May 2013).	128

ACKNOWLEDGEMENTS

In my dissertation defence, I had a slide with a picture of an iceberg. While writing this acknowledgement, I have the same feeling that few lines of acknowledgement is like a tip of an iceberg that does not really reflect the contributions of so many individuals in successfully completing this thesis and my graduate journey.

First of all, I would like to acknowledge my adviser, Prof. Amit Sheth, for his support and guidance throughout my Ph.D. journey. I am greatly thankful to him for letting me be a part of Kno.e.sis Gurukul and for his advice, not only in research but also in many aspects of a successful career. At Kno.e.sis, I got a unique opportunity to work on a variety of the projects and to explore various research areas. I really appreciate the freedom that Dr. Sheth gives to his students in pursuing their research interests yet giving his constructive visionary inputs in aligning the research problems to real-world motivations. I always felt his care for his students professional success and overall well-being. His encouragement greatly helped me to maintain a high motivation through the inevitable ups and downs of my Ph.D. study. Without him, I would not have accomplished what I have right now.

I would like to thank Dr. Jyotishman Pathak for my internship at Mayo Clinic, which helped me to shape my research work. I really appreciate his guidance and support in my research and job search process. Without my Mayo Clinic internship and Jyoti's support, I could not have made it. Furthermore, I am grateful to Dr. T. K. Prasad for his willingness to spend time in understanding my work and providing constructive feedback on my thesis write-up. I always learnt something from my numerous insightful interactions him. I would like to acknowledge Dr. Michael Raymer for his valuable advice, and in-depth discussion during the course of this research. I would also like to acknowledge Dr. Hamid Motahari for giving me the opportunity to do my first research internship

at HP Labs. I have benefited greatly from many interactions with him and really appreciate his continual support and encouragement.

The Kno.e.sis is a great place to pursue research and thrive. It is basically a research eco-system with exciting research projects (from NSF, NIH and AFRL), world-class computing infrastructure, and most importantly inspiring faculty members and awesome students. I feel lucky to be a part of such a great place. At the same time, it is a fun place and a support system (believe me, it is crucial during Ph.D.). I thank all my current and former colleagues for that. I am thankful to Kno.e.sis alumni Drs. Karthik Gomadam, Ajith Ranabahu, Meena Nagarajan, Christopher Thomas, Cartic Ramakrishnan, Pablo Mendes and Satya Sahoo for their guidance during my early stage of Ph.D. and for their willingness to help whenever I asked. I am also grateful to my colleagues Wenbo, Pramod, Hemant, Raghava, Lu, Vinh, Swapnil, Shreyansh, Kalpa, Pavan, Delroy, Sujan, Sarasi, Sanjaya, Harshal, Adarsh, Shiva, Surendra, Nishita, Vaikunth, Ramakant, and Tanvi without them this journey would not be exciting and fun. I wish best to all new kno.e.sis members: Amir, Hussein, Monerie, Utkarshini, and Saeedeh. I also want to thank administrative staff at Kno.e.sis, Tonya Davis, Jibril Ikharo, and John Aguilar for always being so helpful.

Last but not the least, I would like to thank my family, especially my parents and my wife, Rashmi, for constant encouragement, enormous support, and endless love.

This material is based upon work supported by the National Science Foundation under Grant IIS-1111182 SoCS: Collaborative Research: Social Media Enhanced Organizational Sensemaking in Emergency Response. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Dedicated to

my dear parents, Sopan and Sunanda, and my lovely wife Rashmi . . .

1

Introduction

Web search has emerged as a key enabling technology to facilitate faster access to information available on the Internet. It has become an integral part of our lives. Every day, millions of users submit over 3.5 billion of queries to commercial Web search engines such as Google, Yahoo!, and Bing¹. This massive amount of search traffic has boosted the economic underpinning of Web search, namely online advertising, which places relevant advertisements alongside search results by understanding users' search queries. In order to enhance the search experience and improve search ad relevance, Web search is evolving from keyword-based search to semantic search². One of the key aspects in building an intelligent search engine is to understand users' search intents and information needs. Search intent mining can help search engines to enhance their ranking of search results, enable new search features like instant answers, personalization, search result diversification, and the recommendation of more relevant ads. Thus, in the past few years, search intent mining has become a hot topic in Web search and IR research. The intents of a search query can be represented by its search goals [Broder 2002], such as informational, navigational, and transactional. It can also be represented by semantic categories or topics [Sheth et al. 2001; Beitzel et al. 2007]. We define *search intent* as a significant object/topic that denotes users' information needs.

¹<http://www.internetlivestats.com/google-search-statistics/>

²<http://amitsheth.blogspot.com/2014/09/15-years-of-semantic-search-and.html>

Search is shifting towards understanding search intent and serving the appropriate entities. This trend has been driven largely by the increasing amount of structured and semi-structured data, such as relational databases, knowledge sources (i.e. ontologies and Wikipedia), and semantically-annotated Web documents (e.g. schema.org) that have been made available to search engines[Li 2010]. These knowledge sources encode a wealth of information. Searching over such data sources and semantically-annotated documents, in many cases, can offer more relevant and useful results that can satisfy users' information needs. Use of knowledge bases or ontologies for semantic approaches to improving search (as well as browsing, personalization, advertisement) was pioneered around 1990-2002 by Taalee/Semagix [Sheth et al. 2001; Sheth et al. 2002; Hammond et al. 2002]. Recent resurgence of similar approaches that harness knowledge based for search include the Google Knowledge Graph³ and Bing⁴. With the Knowledge Graph advancements, now Google search not only provides a ranked list of relevant web pages but also provides additional important information about searched entities extracted from knowledge bases on the side. For example, a Google search for 'type 2 diabetes' provides essential information such as a description, symptoms, and treatment for type 2 diabetes in a structured format (Figure 1.1). Google search can provide this enhanced Web search experience by understanding users' search intents in terms of semantic entities, linking the entities to domains such as people, health, sports, and movies, and then extracting insightful information about these entities from the relevant knowledge bases.

Although Google's Knowledge Graph is revolutionizing Web search, at present Google search can provide structured faceted information for only few search queries. One major challenge here is understanding search intent, not only at the domain level but within a domain. Understanding the domain of a search query is crucial as it has implications on search result selection and ranking. By understanding the domain of a search query, a search engine can return more relevant and essential results, complimentary structured information, and targeted ads rather than providing keyword-

³<https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>

⁴<https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/>

Google diabetes type 2 Search Query

All News Images Books Videos More Search tools

About 156,000,000 results (0.71 seconds)

Search Ads

Search Results

Information from Knowledge Graph

Type 2 Diabetes Info
www.type2diabetes-information.com/ Get More Information On Type 2 Diabetes & How To Take Care. About Blood Sugar Tips And Advice Important Safety Info Savings Card

Type 2 Diabetes Facts
www.type-2-diabetes-management.com/ Get Info On Treatment Option And Info—for HCPs. Drug Information - Dosing & Administration - Safety Info - Clinical Efficacy

Adult Type 2 Diabetes
Prescription treatment website Learn About A Prescription Option. Find Out About Help For Adult T2D. How It Works - Usage Details - Talk To Your Doctor

Type 2 Diabetes - American Diabetes Association
www.diabetes.org/diabetes-basics/type-2/ American Diabetes Association If you have type 2 diabetes your body does not use insulin properly. This is called insulin resistance. At first, your pancreas makes extra insulin to make up for it. Facts About Type 2 - Where Do I Begin With Type 2? - Treatment & Care

Type 2 Diabetes: Causes, Symptoms, Prevention, and More
www.webmd.com/diabetes/type-2-diabetes.../type-2-diabetes WebMD Most people with diabetes have type 2. What causes this life-long illness? Can you prevent it? How do you know you have it? What can you do about it? Type 2 Diabetes: Causes ... - Type 2 Diabetes in Children - Insulin - Diabetes Testing


In the news

6 Warning Signs Of Type 2 Diabetes That Can Cause Health Complications
Parent Herald - 2 days ago
Type 2 diabetes affects millions of people in the United States causing it to be known as an ...

YMCA of the East Valley offers tips for preventing type 2 diabetes
Redlands Daily Facts - 22 hours ago

Type 2 diabetes
Also called: adult onset diabetes

ABOUT SYMPTOMS TREATMENTS



Testing blood-sugar levels

A chronic condition that affects the way the body processes blood sugar (glucose).

Very common
More than 3 million US cases per year

- Treatable by a medical professional
- Requires a medical diagnosis
- Lab tests or imaging always required
- Chronic: can last for years or be lifelong

With type 2 diabetes, the body either doesn't produce enough insulin, or it resists insulin.
Symptoms include increased thirst, frequent urination, hunger, fatigue,

Figure 1.1: Structured information provided by Google search for a 'type 2 diabetes' query

based search results. While state-of-the-art techniques can identify the domain of the queries (e.g., sports, movies, health), identifying domain-specific intent is still an open problem. Such challenges and domain-specific cognitive systems like IBM Watson Health⁵ have provided an opportunity for advancement and fostered increasing interest in domain-specific search intent mining research.

1.1 Selection of Domain

Among all topics available on the Internet, health is one of the most important in terms of impact on the user and is one of the most frequently searched topics. The Internet is a popular place to learn about health matters. With the growing availability of online health resources, consumers are increasingly using the Internet to seek health-related information [Fox and Duggan 2013; Higgins et al. 2011]. According to a 2013 Pew Survey [Fox and Duggan 2013], one in three American adults has gone online to find information about a specific medical condition. In the current climate of rising health-care costs, the role of freely available and easily accessible health-care information is becoming more central to patients, their families and friends, and even to healthcare providers. Although health information is available in abundance, many Internet users continue to face challenges in accessing relevant, high-quality, and literacy-sensitive health information.

One of the most common ways to seek online health information is via Web search engines such as Google, Bing, and Yahoo!. Approximately 8 in 10 online health inquiries start from a Web search engine [Fox and Duggan 2013]. Non-experts generally lack proper medical knowledge to formulate health search queries by translating their health problems accurately. Search results for health information are often unsatisfactory due to the poor quality input to search engines as well as search engines' failure to understand users' health search intent [Chapman et al. 2003; Keselman et al. 2008; Luo et al. 2008]. Therefore, in spite of the rapid advances in search engine technology, understanding users' health information seeking intents is still challenging. Furthermore, while

⁵<http://www.ibm.com/smarterplanet/us/en/ibmwatson/health/>

working on Mayo Clinic's consumer health information portal, we realised the importance health search intent mining for real-world applications like personalized health information interventions and better understanding of consumers' health information needs. This variety of motivations helped us to envision the broader impact of selecting the health domain for search intent mining on information retrieval and health informatics research with benefits that can be translated to consumers (through the Mayo Clinic portal).

1.2 Knowledge-driven Health Search Intent Mining

This dissertation presents a knowledge-driven approach for domain-specific search intent mining with a focus on health-related search queries. In this study, we have collected health-related search queries originating from search engines that direct users to Mayo Clinic's consumer health information portal (MayoClinic.com). The MayoClinic.com portal is one of the top online health information portals within the United States and on average is visited by millions of unique visitors every day, with around 90% of the incoming traffic originating from Web search engines. This significant traffic to the portal provides us with an excellent platform to conduct our research.

Selection of Intent Classes

To achieve these goals, we first must identify which common intent classes or types of queries are the best abstraction of the users' specific queries. In order to understand users' perspective about online health information seeking, we took a qualitative approach and conducted a focus group study. We studied why, what, and how participants use the Internet to seek health information. Subsequently, we selected 14 consumer-oriented health search intent classes based on:

- Inputs from the focus group study.
- Analysis of health categories on popular health websites (e.g., Mayo Clinic, WebMD).
- A review of health information seeking literature.

- Empirical study of health-related search queries from MayoClinic.com.

Problem statement

Given a set Q of health-related search queries, classify each query q from Q into zero or more intent classes from set IC in a disease-agnostic manner, where IC is a set of 14 consumer-oriented intent classes.

Health domain constraint

There are thousands of health conditions and each health condition has unique characteristics. It is not feasible to develop a separate health search intent mining technique for each health condition. Thus, while developing techniques for health search intent mining it is important that the approach can be generalized and can identify health search intent in a disease-agnostic manner.

Multi-Label classification

As a search query can be classified into zero or more intent classes, the health search intent mining problem is a multi-label classification problem. Unlike binary classification problems, multi-label classification allows the instances to be associated with more than one class. Existing methods for multi-label classification fall into two main categories: a) problem transformation methods (e.g. Binary Relevance [Cherman et al. 2011], Label Power Set [Tsoumakas and Katakis 2006], RAKEL-RANdom k-LabELsets [Tsoumakas and Vlahavas 2007]) and b) algorithm adaptation methods (e.g. tree-based boosting - AdaBoost.MR [Schapire and Singer 2000], ML-kNN [Zhang and Zhou 2007], Rank-SVM [Zhang and Zhou 2014]). Problem transformation methods transform the multi-label classification problem either into one or more single-label classification or regression problems. Algorithm adaptation methods extend specific learning algorithms in order to handle multi-label data directly. Both these methods follow underlying principles of the supervised learning approach and

depend on training data.

Challenges in creation of training data for supervised learning approaches

Training data creation is a manual, costly, and time-consuming process. Depending on the nature of the problem and labeling task, the creation of labeled data for a learning problem often requires domain experts. Moreover, training data suffers from limited coverage (if the training data does not cover all the aspects of the dataset) and a generalization problem. These challenges get amplified for multi-label classification problems as we need to create training data for each label. For our problem, we would be required to create training data for 14 intent classes. Furthermore, we would need domain experts such as healthcare providers and clinicians to label dataset. Moreover, a classifier trained for one disease may not work for other diseases as symptoms, treatments, and medications vary by different diseases. These challenges make supervised learning-based approaches infeasible for solving health search intent mining problem in a disease-agnostic manner.

Biomedical knowledgebases

Over the last decade, biomedical knowledge bases have become an increasingly important component of biomedical research as they encode a vast biomedical knowledge in a structure that can be easily shared and reused by both humans and computers. They contain 1) millions of individual concepts, their meaning and synonyms, 2) relationships between the concepts (e.g. concept hierarchy), and 3) mapping of the concepts to semantic classes. Thus, leveraging rich knowledge from biomedical knowledge sources is an appealing choice for the semantic processing of the health search queries. In this work, we have leveraged rich biomedical knowledge from the Unified Medical Language System (UMLS). UMLS incorporates over 100 medical vocabularies and facilitates computer understanding of biomedical text. Integrated datasets include SNOMED-CT, ICD-x (International Classification of Diseases), NCBI taxonomy, Gene Ontology, the Medical Subject Headings (MeSH), and OMIM. UMLS contains over a million concepts, and maps each concept to Semantic Types

(broader categories, a total 140 types).

Concept identification

The first task in health search intent mining is to identify medical concepts from the search queries. We used UMLS MetaMap for annotating search queries with UMLS Semantic Types and UMLS concepts. In the annotation step, we first addressed the concept identification challenge. While processing multi-word terms, sometimes the MetaMap does not map concepts properly. For example, MetaMap maps the phrase “water in brain” to “water” (Drinking water) and “brain” (brain - body part). The correct mapping of “water in brain” is “hydrocephalus”, which is a medical condition. In order to address this problem, we used advanced text analytics techniques like word sense disambiguation (WSD) and term processing while annotating the data using MetaMap. With the WSD module, MetaMap generates mappings for the terms considering the surrounding text. With the term processing module, MetaMap process each input record as a single phrase in order to identify more complex Metathesaurus terms.

MetaMap data processing

Although MetaMap is a great tool for annotating medical concepts from the search queries, it is very inefficient in terms of scalable data processing. Since the size of our dataset was fairly large (10 million search queries), it was estimated that MetaMap would take a significant amount of time (in days) to annotate 10 million search queries. To address this challenge and to improve data annotation speedup, we developed a scalable MetaMap implementation using a Hadoop-MapReduce framework [Panahiazar et al. 2014]. With this framework, we observed a very significant improvement in the data processing time.

Consumer health vocabulary

Another challenge in solving the health search intent identification problem is the mapping of terms used by laymen to medical terms. Domain experts search for information differently than the people with little or no domain knowledge [White et al. 2009]. Domain expertise is not the same as search expertise since it concerns knowledge of the subject or topic of the information need rather than knowledge of the search process [White and Drucker 2007]. Studies of domain expertise have highlighted several differences between experts and non-experts, including vocabulary and search expression [Allen 1991]. While health domain experts have foundational medical domain knowledge based on formal education and professional experience, laypersons have some socially and culturally derived notions of health and illness acquired from formal and informal sources (e.g., media exposure) and unique personal experiences [Zeng and Tse 2006]. Most of the health search queries are submitted by the laymen (non-experts) and terms used by the laymen are different than the medical terms used by clinicians and healthcare providers. For example, a layman would most likely use “hair loss” to search for information on “alopecia” (the clinical term for hair loss).

Although UMLS contains a Consumer Health Vocabulary (CHV) that maps consumer-driven medical terms to clinical terms, it has limited coverage. For example, for the search query “water on the knee”, even with advanced concept identification techniques and using CHV from the UMLS, the MetaMap maps it incorrectly to “Water thick-knee” (*Burhinus vermiculatus*), which is a bird. “Water on the knee” is actually a consumer-oriented term for a medical condition “knee effusion”. To overcome this challenge, we leveraged knowledge presented in Wikipedia and developed a comprehensive Consumer Health Vocabulary. Wikipedia is the largest and the most visited online encyclopedia. Wikipedia provides complex health information in a simplified way which makes it appealing for both laymen and healthcare professionals. Wikipedia health articles tend to link consumer-oriented terms with health professional's terminology using some semantic relationships (e.g., “Epistaxis, also known as a nosebleed”).

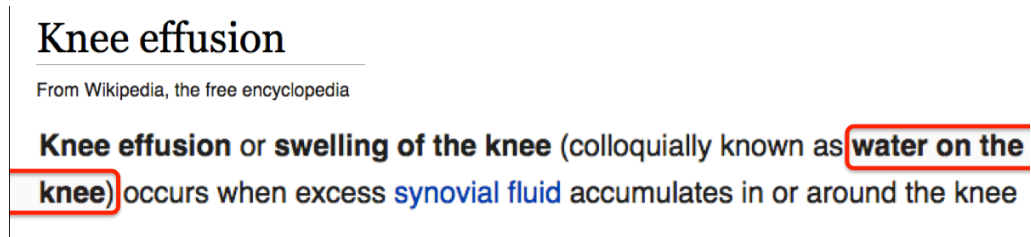


Figure 1.2: a snippet from a Wikipedia article on “knee effusion”.

Motivating Example

Here is a snippet from a Wikipedia article on “Knee effusion” (Figure 1.2). The article mentions alternate terms for “knee effusion”, i.e. “swelling of the knee” and “water on the knee”. This knowledge helps us to map the consumer-oriented term “water on the knee” to the medical term “knee effusion”. Given “knee effusion”, the MetaMap correctly identifies it as a “Disease or Syndrome” concept. Such knowledge makes Wikipedia a very exciting resource for CHV generation. In this research, we exploited such relationships and knowledge from Wikipedia to generate our Consumer Health Vocabulary.

Although using Wikipedia we can generate semantically related candidate term pairs (e.g., knee effusion, water on the knee, swelling of the knee, water on the knee), we cannot identify CHV terms as Wikipedia does not state which term is consumer-oriented and which one is a medical professional term.

Thus, we solved the problem of generating CHV using Wikipedia by addressing the following two subproblems:

- To generate set of candidate pairs from health related Wikipedia articles.
- To identify consumer-oriented terms (CHV term) and health professional medical terms (medical term) from the set of candidate pairs.

We developed a pattern-based information extractor that extracts candidate pairs of CHV and medical terms from health-related Wikipedia pages. We used a hypothesis-based approach to identify

CHV terms. As compared to most of the CHV generation approaches, this approach is automated and does not require manual review of CHV terms from domain experts. Furthermore, it uses knowledge from Wikipedia that is being continuously updated with emerging health terms.

Finally, we developed a semantics-driven search intent mining approach by leveraging rich background knowledge from UMLS and a crowd sourced encyclopedia (Wikipedia) [Jadhav et al. 2014; Jadhav et al. 2014a; Jadhav et al. 2014; Jadhav et al. 2014b]. This approach can identify search intent in a disease-agnostic manner and has been evaluated on the three major chronic diseases: cardiovascular diseases, diabetes, and cancer. In summary, the following are our major contributions in this work:

- Developed an approach to automatically identify health search intents from large-scale search logs in a disease-agnostic manner.
- Constructed a consumer health vocabulary that maps laymen terms to medical terms used by health professionals by parsing health-related Wikipedia articles.
- In the MetaMap data processing, we used advanced text analytics techniques like word sense disambiguation and term processing, and utilized consumer health vocabulary to improve concept identification from the search queries.
- Developed a scalable MetaMap implementation using a Hadoop-MapReduce framework to improve MetaMap's data annotation speedup.

1.3 Twitter, a Health Information Source

While users often turn to search engines to learn about health conditions, a surprising amount of health information is also shared and consumed via social media, such as the public social platform Twitter. Information behavior researchers have described two primary approaches for information acquisition [Lu 2012]. The first is intentional information acquisition, which involves the active

seeking for information and generally triggered by users information needs, e.g., information seeking using Web search. However, in many circumstances users discover information on the Social Web merely by accident (i.e., accidental discovery of information [Erdelez 1997]). This experience of accidental information discovery refers to accidentally bumping into (useful or personal interest-related) information as opposed to intentionally looking for it. Social networking websites such as Facebook and Twitter provide excellent opportunities for accidental information acquisition. In the past few years, Twitter has emerged as one of the major information source that web users are using to keep up with newest health information. A survey [Fox and Jones 2012] indicated that as many as 39% of online health information seekers used social media, and a fraction of them had also followed their contacts' health experiences or updates, posted their own health-related comments, gathered health information, or joined a health-related group. Other research has shown that people prefer search engines while seeking information for various sets of medical conditions, and prefer Twitter for sharing and learning about new health information [De Choudhury et al. 2014].

In some cases people prefer Twitter as an information source, as compared to traditional information sources (e.g. newspapers) [Teevan et al. 2011] since they can find timely information aggregated in one place, information which they would not think to check for on the Internet on their own accord. In many cases, the phenomenon of accidental information discovery is facilitated by users' prior actions. For example, a person who is interested in keeping track of online health information may follow health-related Twitter accounts that can provide him the newest yet reliable health information. This is also known as serendipity [Roberts 1989]; the chance of bumping into unexpected information can be increased by frequently interacting with other people or being exposed to an information-rich environment [Erdelez and Rioux 2000; McCay-Peet and Toms 2010] (here, health-related Twitter accounts). Currently Twitter has thousands of health-centric accounts, which are followed by millions of users to keep up with health information.

Challenges

Although Twitter is an excellent information source, identification of informative tweets from the deluge of tweets is a major challenge. Most of these tweets are highly personal and contextual; hence most of them are neither interesting nor meaningful to anybody outside of the author's circle of followers. In most of the cases, a user has to go through all tweets manually and has to depend on his/her own intellect and analytical capabilities to identify informative tweets. Furthermore, the informativeness of a tweet is subjective and depends upon various factors about the reader, such as the reader's intent, knowledge about the information in the tweet or novelty in the information, interest in the subject, and who authored/shared the tweet (expert in a domain, personal connection).

Thus, to address these problems we have abstracted out the subjective nature of the informativeness problem and objectively defined the tweet informativeness problem. We developed a hybrid approach consisting of rule-based filtering and supervised machine learning for classifying tweets into informative and noninformative categories. In rule-based filtering step, we used following filters: tweets in English language, tweets with URLs, minimum tweet length (5 words and 80 characters) and minimum 5 Google PageRank for URLs. We also filtered-out duplicate tweets, broken and not working URLs. Using the rule-based filtering, we reduced the experiment dataset from 40K tweets to 6.3K tweets (84.25% reduction in the dataset). For the supervised classification, we performed multiple experiments with different classifiers (Naive Bayes, Random Forest, Libsvm). Based on the experiments, we selected a Naive Bayes classifier as it was very fast (a crucial factor for classifying millions of tweets in a timely manner) and had competitive performance with respect to the other classifiers. For the classification, we used following features associated with the tweets and their URLs :

- **Ngrams:** unigrams and bigrams from tweets, URL title and URL content
- **Text features:** length of the tweet, number of special characters, POS tags

1.4. COMPARATIVE ANALYSIS OF EFFECTS OF DEVICE ON EXPRESSION OF SEARCH INTENTS:

- **Author features:** 1) social connectivity, i.e. number of follow-followers, 2) Twitter activity, i.e. number of tweets, and 3) authors credibility or influence, i.e. Klout score
- **Popularity features:** number of retweets, Facebook shares, Facebook likes, Facebook comments, Twitter shares (tweets), and Google Plus shares
- **PageRank:** Google PageRank of the URLs in the tweets

Using Naive Bayes classifier and above mentioned features, we classified 80.93% (precision) of the tweets correctly. Furthermore, we leveraged biomedical domain knowledge to facilitate the retrieval of relevant and reliable health information shared on Twitter in real-time using a system called “*Social Health Signals*” [Soni 2015; Jadhav et al. 2015]. Moreover, to enable efficient browsing of the health information on the Social Health Signals, we are using our search intent mining algorithm which classifies informative tweets and health news into consumer-oriented health categories like Symptoms, Food and Diet, Prevention and Treatments. Such categorization enables users to further filter the informative tweets by health categories of their interest.

1.4 Comparative Analysis of Effects of Device on Expression of Search Intents

So far, we covered topics related to the identification of search intents from the Web search queries and application of intent mining algorithm for Twitter. In the final part of this dissertation (Chapter 6), we compared expression of health search intents and associated features. We performed a large-scale study to compare health search intents and features that contribute in the expression of search intent from more than 100 million search queries from smart devices (smartphones or tablets) and personal computers (desktops or laptops) [Jadhav et al. 2014; Jadhav and Pathak 2014]. In 2015, Google revealed that more Google searches take place on smart devices than on personal computers

in 10 countries, including the US and Japan⁶. With the recent exponential increase in usage of smart devices, the percentage of people using smart devices to search for health information is also growing rapidly [Duggan and Smith 2013], [Fox and Duggan 2012]. Although the user experience for online health information seeking varies with the device used, very few studies have investigated how online health information seeking behavior may differ by device. Understanding the effects of the device used (SDs vs. PCs) for health information search would help us to acquire more insights into online health information seeking behavior. Such knowledge can be applied to improve the search experience and to develop more advanced next-generation knowledge and content delivery systems. To this end, we performed feature-based comparative analysis of more 100 million search queries from PCs and SDs.

1.5 Dissertation Organization

This dissertation is organized as follows: Chapter 2 gives a summary of the past works related to search intent mining. Chapter 3 introduces the problem of health intent mining and discusses techniques used to solve the problem by leveraging knowledge from biomedical domain and Wikipedia. Chapter 4 discusses application of the search intent mining algorithm on health related Twitter data. Since Twitter data is very noisy, we first addressed the problem of identification of informative tweets from noisy Twitter data. Chapter 5 describes the focus group study that we conducted to understand consumers' perspective on online health information seeking and their health search intents. Chapter 6 presents a comparative analysis of health search intents and features that contribute in the expression of search intent from more than 100 million search queries from smart devices (smartphones or tablets) and personal computers (desktops or laptops). Chapter 7 concludes the Dissertation.

⁶<http://adwords.blogspot.com/2015/05/building-for-next-moment.html>

2

Search Intent Mining

Since the last decade, Internet literacy and the number of Internet users have increased exponentially. With the growing availability of online resources, users are increasingly using Web searches to access to information available on the Internet. Everyday, millions of users submit over 3.5 billions of queries to commercial search engines such as Google, Yahoo!, and Bing. In a Web search task a user with an information need describes the information need via a set of query words that are submitted to the Web search engines. Understanding the users latent intents behind the search queries is essential for satisfying a users search needs. It is only through this understanding that search engines will be able to guide the user to obtain the actual desired information. Hence, in recent years, search query intent mining has become one of the important research problems and many approaches have been proposed for mapping search queries into different intent classes. The Merriam-Webster Online Dictionary defines intent as the thing that you plan to do or achieve; an aim or purpose. [Jansen and Booth 2010], define search intent as the expression of an affective, cognitive, or situational goal in an interaction with a Web Search Engine. In this chapter, we will review related work in the search intent mining problem space.

2.1 Background

The search intent mining problem has been stated to have significant overlap with other problems such as search topic mining, search query classification/categorization, search subtopic mining and search goal mining. The intent of a query can be characterized along several dimensions, including search goals [Broder 2002], semantic classes [Beitzel et al. 2007], topics [Beeferman and Berger 2000], and subtopics [Clarke et al. 2009]. The intent of the search queries can also be represented by semantic categories or topics [Broder et al. 2007] [Li et al. 2005] [Pu et al. 2002]. Furthermore, it can be represented by subtopics, denoting multiple senses or multiple facets of the query.

In a pioneering work, Broder [Broder 2002] proposed a search intent taxonomy that is composed of three intents, namely, informational, navigational and transactional. Broder defined these intents as follows:

- Navigational: the immediate intent is to reach a particular site.
- Informational: the intent is to acquire some information assumed to be present on one or more web pages.
- Transactional: the intent is to perform some web-mediated activity.

Broder made a classification of queries through a user survey and manual classification of a query log. In line with Broders work, Rose and Levinson [Rose and Levinson 2004] developed a framework for manual classification of search intents while extending the intent classes proposed by Broder. In their studies Broder, Rose and Levinson showed that the intent of queries can be identified manually.

2.2 Search Intent Mining based on Query Log

Following Broders taxonomy, several authors have focused their work on the automatic classification and characterization of user intents [Jansen et al. 2008] [Lee et al. 2005] [Liu et al. 2006]. Baeza-

Yates et al. [Baeza-Yates et al. 2006] , have worked on large manually annotated data sets, where a data set of around 6,000 popular queries were classified into two aspects: intention (Informational, Not Informational, and Ambiguous) and topic (ODP¹ topics) based on supervised and unsupervised learning techniques. A supervised learning approach is used to identify the user interest given certain established intents and topics; on the other hand, unsupervised learning approach is used to validate the intents and topics used, refine them, and select the one most appropriate to the users needs. Qian et al. [Qian et al. 2013] proposed a method for mining dynamic intents from search query logs. Hu et al.[Hu et al. 2012] proposed a clustering algorithm to automatically mine the subtopics of queries. Dang et al.[Dang et al. 2011] clustered reformulated queries generated from publicly available resources. Sadikov et al.[Sadikov et al. 2010] addressed the problem of clustering the refinements of a user search query. As an orthogonal approach to tackle query intent mining, Li et al.[Li et al. 2008] made the first attempt to increase the amount of training data for query intent mining. Recently the interest in determining user intentions has spread to commercial [Ashkan et al. 2009] and geographical [Gan et al. 2008] applications. For example, in the context of sponsored search, information providers may also wish to know whether a user intends to purchase or utilize a commercial service, or what is called online commercial intention.

2.3 Search Intent Mining for Personalization

Web search personalization (personalization of search results and ranking) is an important area in the field of IR that attempts to tailor search results to a particular user based on that users interests and preferences. In personalization additional context about users, beyond merely the search query issued, is used to enhance rankings, thus providing more effective and efficient information access [Sieg et al. 2007] [Radlinski and Dumais 2006]. One of the most critical factors in Web search personalization is to create a user profile that captures long-term interests. A user interest is generally

¹Open Directory Project <http://www.dmoz.org/>

represented as set of topics searched by a user over a period of time. The problem of the identification of topics from search queries and modeling users interest profiles is considered one of the sub-problem of search intent mining. [Nanda et al. 2014] have used an ontology-based approach for mining users interests and creating user profiles that can be used for Web search personalization. For example, [Nanda et al. 2014] created an ontology-based users interest profiles leveraging topic hierarchy from the Open Directory Project and Wikipedia, combining it with explicit user interests (a users bookmarks, search keywords, and related terms). User profiles are further improved through collaborative filtering using the k-nearest neighbor-based algorithm by terms between similar users.

Ustinovskiy et al. [Ustinovskiy and Serdyukov 2013] considered short-term context (such as queries and click-through data) by exploiting browsing history and search sessions. A search session is a series of intent-related users queries issued to a search engine. Ryen et al. [White et al. 2010] also studied short-term context, current sessions, and queries to predict short-term interests of users by combining and weighing the context of each query. Matthijs and Radlinski [Matthijs and Radlinski 2011] used long-term search history to model users interests in order to re-rank Web results. In the same context, works like Spereta et al. [Speretta and Gauch 2004] proposed user profiling using their search histories. In this work [Harvey et al. 2013] they used query logs to build users topical interest based on the representation of clicked documents over a set of topics determined by latent topic models. Makvana et al. [Makvana and Shah 2014], as opposed to client-side history, analyzed Web logs from servers.

2.4 Search Intent Mining Based on Click-through Data

Click-through data contain the queries submitted by users, followed by the URLs of documents clicked by users for these queries. Click-through data in search engines can be thought of as triplets (q, r, c) consisting of the query q , the ranking r presented to the user, and the set c of links the user clicked on [Joachims 2002]. Lee et al. [Lee et al. 2005] focused on automatic identification

of search intents (navigational and informational) based on the clicks made by the users on the results offered by the search engine. They utilized two major features, users past click behavior and the anchor-link distribution. Click-through bipartite graph data can be used for clustering queries and URLs. Specifically, queries that share the same clicked URLs are considered similar. Methods for performing the task have been proposed e.g. [Beeferman and Berger 2000], [Cao et al. 2008], [Craswell and Szummer 2007], [Fujita et al. 2010], [Jones and Klinkner 2008], [Radlinski et al. 2010], [Wen et al. 2001]. Beeferman et al. [Beeferman and Berger 2000], for example, proposed conducting clustering on a click-through bipartite graph and viewing the obtained clusters as topics covering multiple queries. Radlinski et al. [Radlinski et al. 2010] proposed first using search session data to find similar queries, and then using a click-through bipartite graph to refine the discovered queries that are similar, and finally grouping the similar queries into the same clusters. The clusters containing the same query are then regarded as topics of the query. More recently, Celikyilmaz et al. [Celikyilmaz et al. 2011] proposed a graph summarization algorithm for categorizing a given speech utterance into one of many semantic intent classes. Recently, most commercial search engines provide query suggestions to improve usability. That is, by guessing a users search intent, a search engine suggests queries which may better reflect the users information need. Cao et al. [Cao et al. 2008] used click-through and session data to provide context-aware query suggestions.

2.5 Search Intent Mining for Query Classification

Query classification [Jansen et al. 1998], also referred as query categorization, is classification of user queries into a ranked list of predefined target categories. Such category information can be used to trigger the most appropriate domain (vertical) searches corresponding to a query, search result re-ranking and diversification, and help find the relevant online advertisements. Query classification is different from traditional text classification. Search queries are usually very short and ambiguous, and it is common that a query belongs to multiple categories. Query classification approaches can

be divided into three categories [Cao et al. 2009]. The first category tries to augment the queries with extra data, including the search results returned for a certain query, the information from an existing corpus, or an intermediate taxonomy [Broder et al. 2007] [Shen et al. 2006]. The second category leverages unlabeled data to help improve the accuracy of supervised learning [Beitzel et al. 2005] [Beitzel et al. 2005]. Finally, the third category of approaches expands the training data by automatically labeling some queries in some click-through data via a self-training-like approach [Li et al. 2008].

[Shen et al. 2006] used search engine results as features, including pages, snippets, and titles, and built classifiers based on a document taxonomy. [Broder et al. 2007] transformed the problem of query classification to document classification, which was solved directly in the target taxonomy. Another way to enhance feature representation is the use of word cluster features [Baker and McCallum 1998], [Pereira et al. 1993]. In such an approach, semantically similar words can be grouped into clusters, either by domain knowledge or by statistical methods, and be used as features to improve the generalization performance of a classifier. Similarly, the query classification methods in [Arguello et al. 2009], [Shen et al. 2006] are also based on supervised learning and external knowledge bases are utilized to augment the training features.

2.6 Search Intent Mining for Vertical Selection

Another line of work in search intent mining research is vertical selection. Recently, a number of Web search engines have begun providing access to specialized search services, or verticals, that focus on a specific type of media (e.g., blogs, images, video) or domain (e.g., health, music, travel). The search services have been developed to provide a specialized type of information service to satisfy a users need according to a particular intent [Hu et al. 2009], [Zhou et al. 2012], [Arguello et al. 2010]. Using a more specialized user interface, a vertical search engine can return more relevant and essential results than a general search engine for in-domain web queries. In contrast to prior query

classification and resource selection tasks, vertical selection is associated with unique resources that can inform the classification decision [Arguello et al. 2009].

3

Domain Specific Search Intent Mining

One of the key aspects in building an intelligent search engine is to understand users' search intents and information needs. The IR community has been constantly seeking and advancing techniques to better understand users' search intents and improve their Web search experience. Understanding the domain of a search query is crucial as it has implications on search result selection and ranking. In this chapter, we will presents a knowledge-driven approach for domain-specific search intent mining with a focus on health-related search queries.

3.1 Motivation

3.1.1 Real-world Challenges with Mayo Clinic's Consumer Health Information Portal

This work is motivated by real-world challenges in analyzing incoming search traffic to Mayo Clinic's consumer health information portal (MayoClinic.com). The search traffic consists of search queries originating from Web search engines (such as Google and Bing) that direct users to the May-

oClinic.com portal. The MayoClinic.com portal is one of the top online health information portals within the United States. The portal provides up-to-date, high-quality online health information produced by professional writers and editors. The MayoClinic.com portal is on average visited by millions of unique visitors every day, and around 90% of the incoming traffic originates from Web search engines. Following, are the two primary reasons for initiating domain-specific search intent mining work at the Mayo Clinic.

1. **Better understanding of consumers health information needs**

Mayo Clinic updates the health information presented on the portal periodically based on the consumers' health information needs. Currently, Mayo Clinic utilizes the following approaches for understanding consumers' health information needs:

- **Clues from landing pages:** A landing page is web page on which a user lands or arrives after clicking on the online resources pointing to the landing page. The Mayo Clinic uses a Web analytics tool, IBM NetInsight, to analyze incoming search queries and users information needs based on landing pages.
- **Keyword-based techniques:** The Mayo Clinic uses keyword-based techniques (e.g. keywords such as symptoms, causes, etc.) to understand health information needs.

One of the major limitations of these approaches is that they do not consider the semantics of the search queries. For example, the above approaches can identify a query for “heart attack symptom” as a symptom query, whereas for a search query without explicit mention of a symptom (such as “pain in the left side of chest”), these approaches would fail to map the query to a symptom query. The motivation for this work is to get a better understanding of consumers' health information needs by semantically processing health search queries.

2. **Personalized Health Information Interventions**

Mayo Clinic does health information interventions through emails and health newsletters. eHealth

interventions are of growing importance in the individual management of health and health behaviors [Chan and Kaufman 2011]. Health information or any information is useful for a reader only if the information is relevant to him. Health information intervention can be very beneficial for a patient if he can learn about medical conditions, symptoms, and treatment options that he may need to know about and would not think to check for on the Internet on his own. Such information can be valuable, relevant, and even lifesaving for patients. In order to do targeted information intervention, it is crucial to identify users' interests. Users' health information interests can be of short-term (e.g. seasonal diseases, curiosity for a health condition) or long-term (e.g. chronic diseases, interest in healthy lifestyle). The motivation for this work is to create user interest profiles based on their (both short and long-term) search histories for personalized health information interventions.

3.1.2 Opportunities and Challenges in Health Domain

1. Online Health Information

Among all topics available on the Internet, health is one of the most important in terms of impact on the user and is one of the most frequently searched. In recent years, the quantity and quality of health information available on the Internet has increased substantially. With increased access to reliable, affordable, and high-speed Internet, the percentage of people using the Internet to search and subsequently to learn from online health information is continuously growing. According to a 2013 Pew Survey [Fox and Duggan 2013], one in three American adults has gone online to find information about a specific medical condition. Online health resources are easily accessible and provide information about most of health topics. These resources can help non-experts to make more informed decisions and play a vital role in improving health literacy. In the current climate of the rising costs of health-care, the role of freely available health-care information is becoming more central to patients, their families and friends, and even to healthcare providers.

2. Online Health Information Seeking

One of the most common ways to seek online health information is via Web search engines such as Google, Bing, and Yahoo!. According to the Pew Survey, approximately 8 in 10 online health inquiries start from a Web search engine. Online health information seekers' search queries reflect a wide spectrum of information needs, from specific medical conditions or symptoms, causes and treatments of diseases, to diet information to healthy lifestyle tips ([Bessell et al. 2002; Nicholas et al. 2003; Andreassen et al. 2007; Zhang and Fu 2011]). Aside from trying to learn more about a symptom or disorder specifically relevant to the person searching, half of online health information research is on behalf of a friend or relative [Sadasivam et al. 2013].

3. Challenges in Online Health Information Seeking

Although health information is available in abundance, many Internet users continue to face challenges in accessing relevant, high quality, and literacy-sensitive health information [Bodie and Dutta 2008; Knapp et al. 2011; Bonnar-Kidd et al. 2009; Connolly and Crosby 2014]. Health literacy is defined as the degree to which individuals can obtain, process, and understand the basic health information and services needed to make appropriate health decisions [Nielsen-Bohlman et al. 2004]. Non-experts generally lack proper medical knowledge to formulate health search queries by translating their health problems accurately. Search results for health information are often unsatisfactory due to the poor quality input to search engines as well as search engines' failure to understand users' health search intent ([Chapman et al. 2003; Keselman et al. 2008; Luo et al. 2008] even go as far as to describe searching for health information as a “trial-and-error” process. Other studies have suggested that search engines should specifically optimize for health search queries [Berland et al. 2001; Benigeri and Pluye 2003]. Therefore, in spite of the rapid advances in search engine technology, understanding users' health information seeking intents in the specialized domain of health information is still challenging.

This variety of motivations helped us to envision the broader impact of selecting the health

domain for search intent mining on information retrieval and health informatics research with benefits that can be translated to consumers (through the Mayo Clinic portal).

3.2 Health Search Intent

- **Definition**

Health information search intent can be interpreted as:

- **Search goals** such as diagnosis (e.g. diagnostic search based on symptoms or health conditions) and learning and exploration.
- **Search topics** such as symptoms, treatments, and prevention.

In this work, we define health search intent as a significant health topic that denotes consumers' health information needs.

One important aspect of this definition is the focus on consumers' health information needs. Here, the consumer refers to all the people that are using the Internet for health information seeking, which constitutes non-experts as well as experts with medical knowledge. Since the percentage of medical experts (healthcare providers and clinicians) is significantly less than the percentage non-experts, researchers have considered consumers as non-experts (laymen).

- **Constraint**

There are thousands of health conditions and each health condition has unique characteristics. It is not feasible to develop a separate health search intent mining technique for each health condition. Thus, while developing techniques for health search intent mining it is important that the approach can be generalized and can identify health search intent in a disease agnostic manner.

- **Objective**

Our defined objective in this work is “to identify consumer-oriented health search intents (topics) from health search queries in a disease agnostic manner”.

3.3 Selection of Health Search Intent Classes

How is one to define a semantic representation that can precisely understand and distinguish the intent of the input query? In this research, we referred to health search intent classes as consumer-oriented health topics that are easily understandable for a non-expert, lay population. Although there are multiple websites, blogs, and forums dedicated to consumer-oriented health content, there is no standardized list of consumer-oriented health topics. Also, most of the medical vocabularies, ontologies, and taxonomies are developed from the perspective of clinicians and health providers. Moreover, even though both IR and health informatics communities have been studying the online health information seeking phenomena, there is a dearth of work on formalizing consumers' health search intents. To address these challenges, we 1) first took qualitative approach and conducted a focus group study to understand consumers' perspective about online health information seeking, 2) analyzed health categories on popular health websites (e.g. Mayo Clinic, WebMD), 3) reviewed health information seeking literature, and 4) empirically studied health-related search queries from MayoClinic.com.

3.3.1 Online Health Information Searching: A Qualitative Approach for Exploring Consumer Perspectives

This is a brief summary of this study. We will cover the study in detail in chapter 5.

1. Background

The Internet is a common resource that patients and consumers use to access health-related information. Multiple practical, cultural, and socioeconomic factors influence why, when, and how people utilize this tool. Improving the delivery of health-related information necessitates a thor-

ough understanding of users' searching-related needs, preferences, and experiences. Although a wide body of quantitative research examining search behavior exists, qualitative approaches have been under-utilized and provide unique perspectives that may prove useful in improving the delivery of health information over the Internet.

2. Objective

We conducted this study to gain a deeper understanding of online health-searching behavior in order to inform future developments of personalizing information searching and content delivery.

3. Approach

We completed three focus groups with adult residents of Olmsted County, Minnesota, which explored perceptions of online health information searching. Participants were recruited through flyers and classified advertisements posted throughout the community. We audio-recorded and transcribed all focus groups, and analyzed data using standard qualitative methods. The study focused on four major aspects:

- (a) Participants' perception and understanding of healthcare information.
- (b) The process of health information search and frequently searched health topics.
- (c) Understanding and usage of information.
- (d) Implications of healthcare information for their health and well-being.

4. Results

Almost all participants reported using the Internet to gather health information. They described a common experience of searching, filtering, and comparing results in order to obtain information relevant to their intended search target. We also collected information about the type of health topics that they search for online. Information saturation and fatigue were cited as the main reasons for terminating searching. This information was often used as a resource to enhance their interactions with healthcare providers.

5. Conclusion

Many participants viewed the Internet as a valuable tool for finding health information in order to support their existing health care resources. Although the Internet is a preferred source of health information, challenges persist in streamlining the search process. Content providers should continue to develop new strategies and technologies aimed at accommodating diverse populations, vocabularies, and health information needs.

This study provided important insights and helped us to understand:

- Consumers' perspective (e.g. their experiences, challenges) about online health information seeking.
- Why (motivations) and how (search strategies) participants use the Internet to seek for health information.
- What health information do they search using the Internet.

3.3.2 Analysis of Health Categories on Popular Websites

The critical factors in selecting consumer-oriented health information provider websites are that the website should be popular among consumers and it should provide high-quality information that is vetted by experts. In order to select such websites, we utilized Google PageRank, Alexa ranking, and ranking from Medical Library Association (CAPHIS).

- **Google PageRank:** It is an algorithm used by Google's search engine to rank websites in Web search results. The PageRank values range from 0 to 10, with higher values indicating greater importance. Google search uses more than 200 signals to calculate a website's PageRank, which indicates its overall importance, authority, and reliability.
- **Alexa Ranking:** Alexa provides traffic data, global rankings, and other information on 30 million websites. Alexa traffic rank is a measure of the website's popularity. It is based on

three months of aggregated historical traffic data from millions of users and data obtained from diverse traffic data sources. It is a combined measure of page views and users. The website also provides ranked list of top websites by topics such as health, music, news, and weather.

- **CAPHIS Ranking:** The Consumer and Patient Health Information Section (CAPHIS) connects health sciences librarians and other consumer health information specialists with a forum. CAPHIS also provides a ranked list of online consumer-oriented websites based on content, credibility, up-to-date information, and several other factors.

Finally, by combining above three ranking, we selected the following websites: MedlinePlus¹, Mayo Clinic², WebMD³, CDC⁴, HealthFinder.gov⁵, and Familydoctor.org⁶. For the selected websites, we studied health topics used for health content organization. Some of these websites have some overlapping health categories while some categories are different (Table 3.1). Difference in the categories is generally due to different way of grouping of the health topics.

3.3.3 Survey of Health Information Seeking Literature

We studied health information seeking literature which spans across more than two decades and multiple disciplines such as computer science (IR, human computer interaction, semantic web), health informatics, and sociology. The literature review helped us to understand, how researchers have sliced and diced health search queries while working on research problems. Apart from popular health topics such as diseases, symptom, cause, and treatment, researchers have also considered other health topics such as information seeking for different age-groups, wellness, disease management, and diet.

¹<http://www.nlm.nih.gov/medlineplus/>

²<http://www.mayoclinic.org/>

³<http://www.webmd.com/>

⁴<http://www.cdc.gov/>

⁵<http://healthfinder.gov/>

⁶<http://familydoctor.org/familydoctor/en.html>

Mayo Clinic	WebMd	MedlinePlus	FamilyDoctor.org
Symptom	Symptom and cause	Symptoms	Symptoms
Cause	Diagnosis & Tests	Diagnosis and Tests	Causes & Risk Factors
Risk	Treatments	Living with	Diagnosis & Tests
Complications	Living with	Treatments and Therapies	Treatment
Test and diagnosis	Complications (Risk)	Related Issues	Complications
Treatments and drugs	Drug and supplements	Disorders and Conditions	Prevention
Lifestyle		Demographic Groups	
Prevention			

Table 3.1: List of health categories on popular health websites

3.3.4 Empirical Study of Health Queries

In this work, we used incoming health search queries from Mayo Clinics consumer health information portal. We created two sets: first, one with the top 100 search queries (based on number of users who submitted the same query in a months time); second, one with 100 randomly selected search queries from a period of a month. We manually studied the search queries from both sets and identified emerging health topics.

Finally, we compiled a list of 14 consumer-oriented intent classes based on the inputs from the focus group study, analysis of health topics on the popular health websites, review of the health information seeking literature, and empirical study of the health search queries. Although, we found this list as representative list of major consumer-oriented intent classes, we do not claim that this list is comprehensive. Note that there can be possible overlaps between some of the intent classes, for example, in a broader sense Drugs and Medications can be considered as a part of Treatment, but in our analysis we considered both as separate intent classes in order to study search traffic for

Intent Classes	Description and Examples
Symptoms	Queries for signs and symptoms, e.g., stroke symptoms, heart palpitations with headache, home remedies for heart murmur, heartburn vs heart attack symptoms.
Causes	Queries related to cause/reasons for various CVD conditions and symptoms, e.g., causes of an elevated heart rate, heart failure reasons, and morning hypertension causes.
Risks and Complication	Queries related to risk and complications, e.g., risks of pacemaker, risk factors to hypertension, complications of bypass surgery, and heart ablation surgery risks.
Drugs and Medications	Queries related to drugs and medications, e.g., dextromethorphan blood pressure, medications hypertension, tylenol raise blood pressure, and ibuprofen heart rate.
Treatments	Queries related to treatments, e.g., exercise for reducing hypertension, cardiac arrest treatments, bypass surgery, and cardiac rehabilitation.
Tests and Diagnosis	Queries related to tests and diagnosis, e.g., heart echocardiogram, diagnosis of vascular disease, ct scan for heart, test for cardiomyopathy, and urinalysis.
Food and Diet	Queries related to food and diet, e.g., what is cardiac diet, what foods lower blood pressure and cholesterol, red wine heart disease, alcohol and hypertension
Living with	To control, management, curing and living with CVD, e.g., exercises to lower high blood pressure, cure for postural hypotension, lifestyle changes to lower hypertension, and how to control cholesterol.

each intent type individually. These intent classes and the classification scheme (Table 3.3.4) are reviewed and verified by the Mayo Clinic clinicians and domain experts.

Intent Classes	Description and Examples
Prevention	Queries related to prevention, e.g., ways to prevent heart attack, foods to avoid heart diseases, aspirin for prevention of stroke, and foods to lower risk of heart disease.
Side effects	Search queries related to side effects, e.g., blood pressure pills side effects, side effects of beta blockers for hypertension, and coq10 bp side effects
Medical devices	Queries related to medical device references, e.g., living with a pacemaker, using blood pressure cuff, pump for pulmonary hypertension, and blood pressure monitor.
Diseases & conditions	Queries related to diseases and conditions, e.g., born with holes in heart, stroke tia symptoms, hypotension, and heart attack in pregnancy.
Age-group References	Queries related to age groups, e.g., cardiac defects in children, average heart rate for an adult, hypertension in adolescents, and heart murmurs in infants.
Vital signs	Queries with references to blood pressure, heart rate, pulse rate, temperature, heart beat (w/o high/low blood pressure as we considered them under (Diseases and Conditions), e.g., blood pressure 125/90, normal resting heart rate, can tylenol raise blood pressure, and healthy heart rate chart

Table 3.2: List of health intent classes and their description with examples

3.4 Problem Statement

Let

- Q be a set of health related search query,
- IC be a set of consumer-oriented intent classes, and
- q be a search query such as $q \in Q$

Classify each query q from Q into zero or more intent classes from set IC , in a disease agnostic manner.

It is a multi-label classification problem.

3.5 Multi-label Classification

Query classification, also referred as query categorization, is classification of user search queries into a list of predefined target classes. Query classification is different from traditional text classification. Search queries are usually very short and ambiguous, and it is common that a query belongs to multiple categories. Query classification problems are generally solved using supervised learning methods. In supervised learning, a model is learned using a set of fully labeled items, which constitute the training set. Once a model is learned, it can be applied to a set of unlabeled items, called the test set, in order to automatically apply labels. One fundamental assumption adopted by traditional supervised learning is that each item can only have one label. Although traditional supervised learning is prevailing and successful, there are many learning tasks where the above simplifying assumption does not fit well as real-world objects might be complicated and have multiple meanings simultaneously [Zhang and Zhou 2014]. To account for the multiple meanings that one real-world object might have, one direct solution is to assign a set of proper labels to the object to explicitly express its semantics. Following the above consideration, the paradigm of multi-label

learning naturally emerges [Zhang and Zhou 2014].

Single label (binary) classification is a common learning problem where the goal is to learn from a set of instances, each associated with a unique class label from a set of disjoint class labels L . Depending on the total number of disjoint classes in L , the problem can be identified as binary classification (*when* $|L| = 2$) or multi-class classification (*when* $|L| > 2$) problem. Unlike binary classification problems, multi-label classification allows the instances to be associated with more than one class. That is, the goal in multi-label classification is to learn from a set of instances where each instance belongs to one or more classes in L . For example, in in-text classification, a news article may include multiple topics such as politics, economics, and health. Similarly, a health search query can be classified into multiple intent classes, e.g., “red wine to control heart disease” may fall into the “Food and Diet”, “Healthy Living”, and “Diseases and conditions” intent classes.

Existing methods for multi-label classification fall into two main categories: a) problem transformation methods [Tsoumakas and Katakis 2006], and b) algorithm adaptation methods. Problem transformation methods transform the multi-label classification problem either into one or more single-label classification or regression problems. Algorithm adaptation methods extend specific learning algorithms in order to handle multi-label data directly. Briefly, the key philosophy of problem transformation methods is to *fit data to algorithm*, while the key philosophy of algorithm adaptation methods is to *fit algorithm to data* [Zhang and Zhou 2014].

3.5.1 Problem Transformation Methods

Problem transformation methods map the multi-label classification task into one or more single-label classification or regression tasks. The baseline approach, called the Binary Relevance [Tsoumakas and Katakis 2006; Cherman et al. 2011] method, decomposes the multi-label classification problem into several independent binary classification problems, one for each label which participates in the multi-label problem. The final multi-label prediction for a new instance is determined by aggregating the classification results from all independent binary classifiers. In recent years, many

approaches have been proposed to further improve classification performance by incorporating the label correlations [Cheng et al. 2010; Hariharan et al. 2010] or exploiting the label hierarchy [Bi and Kwok 2011]. Although these methods can be very accurate on small datasets, they are very slow or even intractable on larger datasets, like [Fürnkranz et al. 2008] and [Cheng et al. 2010]. This necessarily restricts their usefulness since many multi-label contexts involve large numbers of examples and labels.

In a problem transformation method, called Label Power Set (LP) [Tsoumakas and Katakis 2006; Cherman et al. 2011], the multi-label problem can be transformed into one multi-class single-label learning problem, using target values for the class attribute and all unique existing subsets of multi-labels present in the training instances (the distinct subsets of labels). The main drawback of this approach is that the number of label combinations grows exponentially with the number of labels. For example, a multi-label data set with 10 labels can have up to $2^{10} = 1024$ label combinations. This increases the runtime of classification and is not suitable for problems with more labels. The RAKEL (RANdom k-LabELsets) [Tsoumakas and Vlahavas 2007] algorithm iteratively constructs an ensemble of m Label LPclassifiers, each trained on a random subset of the actual labels. Prediction using this ensemble method proceeds by a voting scheme [Tsoumakas and Vlahavas 2007; Cherman et al. 2011]. Classifier chains are an alternative ensembling methods used in multi-label classification. The Calibrated Label Ranking [Fürnkranz et al. 2008] approach transforms the task of multi-label learning into the task of label ranking.

Any single-label learning algorithm can be used to generate the classifiers used by the problem transformation methods. While addressing multi-label classification problem using a problem transformation method, previous work has used Support Vector Machines [Godbole and Sarawagi 2004], Naive Bayes [McCallum 1999], k Nearest Neighbor methods [Spyromitros et al. 2008], and Perceptrons [Fürnkranz et al. 2008] for signal-label classification tasks.

3.5.2 Algorithm Adaptation Methods

The focus of the algorithm adaptation approach aims to tackle multi-label learning problem by modifying existing algorithms so that they can deal with multi-label data directly, without requiring any preprocessing. Well-known approaches include AdaBoost, decision trees, and lazy methods. Such methods are usually chosen to work specifically in certain domains, for example, decision trees are especially popular in bioinformatics. Some adaptations involve problem transformations internally which may be generalizable.

- **Tree-based Boosting:** AdaBoost.MH and AdaBoost.MR [Schapire and Singer 2000] are two simple extensions of AdaBoost for multi-label data where the former tries to minimize Hamming loss and the latter tries to find a hypothesis with optimal ranking. Furthermore, ADABOOST.MH can also be combined with an algorithm for producing alternating decision trees [De Comité et al. 2003]. The resulting multi-label models of this combination can be interpreted by humans.
- **Lazy Learning:** There are several lazy learning-based approaches (i.e., the k Nearest Neighborhood (kNN)) that use either problem transformation or algorithm adaptation [Zhang and Zhou 2007; Wieczorkowska et al. 2006; Brinker and Hüllermeier 2007]. The ML-kNN algorithm extends the k-NN classifier to multi-label data. The basic idea of this algorithm is to adapt k-nearest neighbor techniques to deal with multi-label data where a maximum a posteriori (MAP) rule is utilized to make prediction by reasoning with the labeling information embodied in the neighbors [Zhang and Zhou 2007; 2014].
- **Ranking Support Vector Machine (Rank-SVM):** One important problem with tree-based boosting [Schapire and Singer 2000] is that, they are likely to overfit with relatively smaller (<1,000) training set. Elisseff et. al. in [Elisseff and Weston 2001] proposed a ranking approach for multi-label learning that is based on SVMs algorithm that has an intuitive way

of controlling such complexity while having a small empirical error. The basic idea of this algorithm is to adapt a maximum margin strategy to deal with multi-label data, where a set of linear classifiers are optimized to minimize the empirical ranking loss and enabled to handle nonlinear cases with kernel tricks [Zhang and Zhou 2014]

- **Neural Network:** Neural Networks and Multi-layer perceptron-based algorithms are also have been extended for multi-label data. In BP-MLL [Zhang and Zhou 2006], the error function for the very common neural network learning algorithm, back-propagation has been modified to account for multi-label data.
- **Decision Trees:** Multi-Label C4.5 (ML-C4.5) [Clare and King 2001] is an adaptation of the well-known C4.5 algorithm. The learning process is accomplished by allowing multiple labels in the leaves of the tree, the formula for calculating entropy is modified for solving multi-label problems.

3.5.3 Challenges and Limitations

Following are some of the key challenges in utilizing supervised learning-based multi-label classification approach for solving health intent mining problem with 14 intent classes (i.e. multi-label classification problem with 14 labels).

Following are some key challenges associated with training data:

- **Challenges in training data generation:**
 - **Manual process:** Creation of training data is a manual process in which human annotators label a set of instances from the experiment dataset with the appropriate class label. This is time consuming and labor intensive process.
 - **May require domain experts:** Depending on the nature of the problem and labeling task, the creation of labeled data for a learning problem often requires domain experts. Training data creation with the help of a domain expert is very expensive.

- For our problem, we need annotators with medical knowledge, such as healthcare providers and clinicians, to label training data.

- **Limited coverage:** Ideally training data should be a representative sample of the entire dataset. But in the real world, it is very difficult to create a training dataset that can cover all aspects (discriminative features) of the dataset, and if the training data does not cover all the aspects of the dataset the model learned from such training data often performs poorly on unseen data. This is also known as a generalization problem. Recall that generalization refers to the ability to produce correct outputs for inputs not encountered during the training.

These challenges get amplified for multi-label classification problems, as we need to create training data for each label. For our problem, we would be required to create training data for 14 intent classes. Furthermore, we would need domain experts such as healthcare providers and clinicians to label dataset. Moreover, a classifier trained for one disease may not work for other diseases as symptoms, treatments, and medications vary by different diseases.

These challenges make supervised learning-based approaches infeasible for solving health search intent mining problem in a disease agnostic manner.

3.6 Knowledge-driven Approach

Knowledge bases such as dictionaries, taxonomies, and ontologies encode a wealth of information. These knowledge bases facilitate representation of the knowledge that could be machine-processable, used, and shared among distributed applications and agents. Being machine readable and constructed from the consensus of a community of users or domain experts, they represent a very reliable and structured knowledge source. Such world knowledge in turn enables cognitive applications and knowledge-centric services like disambiguating natural-language text, entity linking, question-answering, and semantic search over entities and relations in Web data. Prominent examples of how knowledge bases can be harnessed for real-world applications include the Google Knowledge

Graph and the IBM Watson question-answering system [Hoffart et al. 2015]. In fact, comprehensive knowledge bases in machine-readable representations have been an elusive goal of AI for decades.

A paradigmatic example is WordNet (Fellbaum, 1998), a domain-independent, and general-purpose thesaurus that describes and organizes more than 117,000 general English concepts, which are semantically structured in an ontological fashion. It contains words (nouns, verbs, adjectives, and adverbs) that are linked to sets of cognitive synonyms (synsets), each expressing a distinct concept (i.e., a word sense). Synsets are linked by means of conceptual semantic and lexical relations such as synonymy, hypernymy (is-a), six types of meronymy (part-of), antonymy, complementary, and so on. The result is a network of meaningfully related words, where the graph model can be exploited to interpret the semantics of the concept. Semantics refers to the meaning of a concept in a context, as opposed to its form (syntax). WordNet has been extensively used as the background knowledge in multiple text processing applications such as word sense disambiguation, question-answering and information retrieval (to expand both queries and document indexing entries).

Pioneering work in knowledge-driven search system is done by Sheth et al. [Sheth et al. 2001; Sheth et al. 2002] in early 2000. Sheth et al. developed comprehensive ontology covering over 25 domains such as sports, entertainment, and news. Sheth implemented automated intelligent agents that can extract meaningful information and metadata from variety of input sources in a structured format. The extracted information is further used to construct knowledge-base. Application of this system includes semantic search and personalization. In the last few years, knowledge bases have evolved from human-created to machine-created ones. The great success of Wikipedia and algorithmic advances in information extraction have enabled the automated or semi-automated creation of large-scale knowledge bases. Recent endeavors of this kind include academic research projects such as DBpedia, KnowItAll, ReadTheWeb, and YAGO, as well as industrial ones such as Freebase, Google Knowledge Graph, Amazon's Evi, and Microsoft's Satori [Hoffart et al. 2015].

3.6.1 Biomedical Knowledge Bases

Over the last decade, biomedical knowledge bases have become an increasingly important component of biomedical research as they encode vast biomedical knowledge in a structured format that can be easily shared and reused by both humans and computers. They contain many millions of individual entities, their mappings into semantic classes, and relationships between entities.

Several biomedical knowledge sources are available freely. Following are some examples.

1. Unified Medical Language System (UMLS)

The National Library of Medicine (NLM) produces the Unified Medical Language System (UMLS) to facilitate computer understanding of biomedical text. The UMLS is a repository of more than 100 biomedical vocabularies. Integrated datasets include SNOMED-CT, ICD-X (International Classification of Diseases), NCBI taxonomy, Gene Ontology, the Medical Subject Headings (MeSH), OMIM, etc. The UMLS consists of three subcomponents.

- **Metathesaurus**⁷

The Metathesaurus forms the base of the UMLS and comprises over 1 million biomedical concepts and 5 million concept names, all of which stem from the over 100 incorporated controlled vocabularies and classification systems. It contains information about biomedical and health related concepts and the relationships among them. Each concept is an abstract representation of the term phrases, which are considered as synonymous in the medical domain. In the Metathesaurus, each concept is given a unique identifier, and all synonymous concepts have the same identifier. This feature helps NLP systems to cluster equivalent terms into unique concepts. It links alternative names and views of the same concept from different source vocabularies.

- **Semantic Network**⁸

⁷<https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html>

⁸<http://semanticnetwork.nlm.nih.gov/>

The Semantic Network consists of Semantic Types and Semantic Relationships. Semantic Types are broad subject categories like Disease or Syndrome and Clinical Drug. Semantic Relationships are useful relationships that exist between Semantic Types. Each concept in the Metathesaurus is assigned one or more Semantic Types (categories), which are linked with one another through Semantic Relationships. The Semantic Network is a catalog of these Semantic Types and Relationships. This is a rather broad classification; there are 135 Semantic Types, and 54 Relationships in total.

- **SPECIALIST Lexicon**⁹

The SPECIALIST lexicon is an English-language lexicon that contains biomedical terms. The lexicon entry for each word or term records the syntactic, morphological, and orthographic information of the respective lemma. It also contains spelling variants, acronyms, and abbreviations.

2. **SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms)**¹⁰

SNOMED CT is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED CT is considered to be the most comprehensive multilingual clinical healthcare terminology in the world. It is developed for clinical decision support, improved patient safety and knowledge-based access to health information in support of the clinical practice of medicine. It is essentially the sets of concepts with each concept designated by a unique identifier and described by terms and hierarchical relationships.

3. **MEDLINE**¹¹ and **PubMed**¹²

MEDLINE is a comprehensive online database of biomedical literature maintained through the National Library of Medicine (NLM). It is the largest and most widely used biomedical

⁹<https://www.nlm.nih.gov/pubs/factsheets/umlslex.html>

¹⁰<https://www.nlm.nih.gov/healthit/snomedct/index.html>

¹¹<https://www.nlm.nih.gov/bsd/pmresources.html>

¹²<http://www.ncbi.nlm.nih.gov/pubmed>

bibliographic database in the world. MEDLINE documents are currently indexed by human experts based on a controlled list of indexing terms derived from the Medical Subject Headings (MeSH) vocabulary. MEDLINE can be accessed via a search engine called PubMed. PubMed offers several tools that help the user define a medical search.

4. Medical Subject Headings (MeSH)¹³

MeSH is the National Library of Medicine's controlled vocabulary of terms used for indexing PubMed articles. MeSH terms are added to bibliographic citations during the process of MEDLINER indexing. MeSH terms constitute a thesaurus that embodies all the concepts appearing in the medical literature. It consists of sets of terms naming descriptors in a hierarchical structure (13-level hierarchy) that permits searching at various levels of specificity using MeSH headings and subheadings. All scientific articles are indexed using an average of 10 to 12 descriptive MeSH terms.

These knowledge bases provide essential domain knowledge to the drive following classes of biomedical applications:

- Search and query of heterogeneous biomedical data
- Data exchange among applications
- Describing biological entities and relationships
- Data annotation
- Information integration
- Natural Language Processing (e.g., relation extraction, document summarization, question-answering, and literature-based discovery)
- Computer reasoning with data

¹³<https://www.nlm.nih.gov/mesh/>

- Information retrieval

3.6.2 In the Context of Health Search Intent Mining

One important aspect of search intent mining is to understand the semantics of the query terms. As mentioned in earlier sections, biomedical knowledge sources encode rich biomedical knowledge in structured and machine processable format. They consist of:

- Concepts, their meaning and synonyms.
- Mapping of concepts to their alternate forms and concepts in other vocabularies.
- Concepts spelling variants, acronyms and abbreviations.
- Relationships between concepts (concept hierarchy) and 54 types of Semantic Relationships.
- Mapping of the concepts to broad subject categories, i.e., to 135 types of Semantic Types.

Thus, leveraging rich knowledge from biomedical knowledge sources is a natural choice for semantic processing of the health search queries. In this work, we have utilized the UMLS as a knowledge base.

3.7 Concept Identification

The first task in our knowledge-driven approach for health search intent mining is to identify medical concepts from the search queries.

Identifying medical concepts from text is one of the major research topics both in Natural Language Processing and biomedical text mining that has spurred the development of several toolkits [Aronson and Lang 2010] [6] such as MetaMap and cTakes. Concept identification, also known as term identification, aims at the identification of meaningful linguistic expressions. In the UMLS Glossary¹⁴, a term is defined as: “A word or collection of words comprising an expression”. In the

¹⁴https://www.nlm.nih.gov/research/umls/new_users/glossary.html

Metathesaurus, a term is the class of all strings that are lexical variants (made singular and normalized to case) of each other. The process of concept identification consists of two primary tasks: concept recognition, and concept mapping.

Example: “what are the medications for stomach pain?”

Concepts: medication, stomach pain.

Concept identification is a challenging task. Following are some of the challenges:

- Lexical or orthographic variants, e.g., diet and dieting and ICD9 and ICD-9.
- Misspelling, e.g., pneumonia: neumonia.
- Synonyms, e.g., heart attack: myocardial infarction.
- Abbreviations, e.g., myocardial infarction: MI.
- Identifying concept boundary (Named Entity Recognition, e.g., pain in stomach= stomach pain)
- Contextual meanings, e.g., “discharge from hospital” versus “discharge from wound”.
- Ambiguous relations among words, e.g., “no acute infiltrate”, which could mean that there is no infiltrate or that there is an infiltrate, but it is not acute.

Lexicon-based approach, rule-based approach, and statistical machine learning-based approach are the popular techniques used in the concept identification task. Linguistic approaches are mainly used to identify phrases that, based on their syntactic form, can serve as candidate terms. Statistical approaches are used to measure the term-hood of phrases. In many cases, linguistic, rule-based, and statistical ML approaches are combined in a single hybrid approach. The UMLS Metathesaurus has also been commonly used as a lexicon for medical text. In the concept-mapping task, the terms are linked to a reference vocabulary. Concept mapping is only possible using lexicon-based concept identification approach.

3.7.1 Medical Concept Identification Tools

- **UMLS MetaMap**

MetaMap is developed by the NLM with the aim to provide better access to biomedical text by extracting entities relevant to the biomedical domain. MetaMap identifies Metathesaurus concepts in free-form textual input and maps them into concepts from the Unified Medical Language System (UMLS) Metathesaurus. The current open-source release consists of following components: word sense disambiguation, lexical and syntactical analysis, variant generation, and POS tagging. MetaMap has been widely used to process datasets ranging from health search queries [Dogan et al. 2009; Herskovic et al. 2007] to emails 15 [Brennan and Aronson 2003] to clinical records. 6 [Aronson and Lang 2010] Concept identification is realized by dictionary lookup. The resulting annotations are provided as mappings to the UMLS Metathesaurus concepts, together with a score that incorporates aspects of centrality, variation, coverage, and cohesiveness.

- **cTAKES**

The Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) [Savova et al. 2010] is a highly modular software system that enables information extraction from electronic medical records and clinical free-text. The cTAKES is built on existing open-source technologies, UIMA and OpenNLP natural language processing toolkit. Modules cover aspects such as text pre-processing, sentence splitting, and tokenization, but also more complex algorithms like negation (based on NegEx [Chapman et al. 2001]) and NERthe latter trained on Mayo Clinic EHRs. Its components are specifically trained for the clinical domain and it provides rich linguistic and semantic annotations [Savova et al. 2010].

- **NCBO Annotator (NCBO)**

The National Center for Biomedical Ontology (NCBO) Annotator [Jonquet et al. 2009] (formerly referred to as the Open Biomedical Annotator (OBA)) [5] annotates text with ontological con-

cepts from all the ontologies contained in the NCBO BioPortal and the UMLS Metathesaurus. During the first stage, NCBO annotator assigns annotations to the textual content based on linguistic features. While in the second stage, these annotations are enriched based on semantic features. The Annotator uses Mgrep2 [Dai et al. 2008] to recognize concepts by using string matching on the dictionary. Mgrep applies stemming as well as permutations of the word order combined with a radix-tree-search algorithm to allow for the identification of the best matches of dictionary entries to a particular text span.

- **MedLEE**

Columbia University's proprietary Medical Language Extraction and Encoding System (MedLEE) was designed for decision support applications in the domain of radiology to process x-ray reports. Later it was extended to other parts of the medical field. MedLEE also extracts a series of modifiers linked to concepts, such as certainty, status, location, quantity, and degree. Applicable concepts are further encoded to the UMLS Metathesaurus concepts.

In this work, we have used UMLS MetaMap for identifying medical concepts from the health search queries.

3.7.2 Concept Identification using MetaMap

The MetaMap first breaks the text into phrases and then, for each phrase, it returns the mapping options ranked according to the strength of the mapping.

Following are some of the lexical/syntactic analysis components within MetaMap that process input text:

- Tokenization: Breaks input text into noun phrases
- Acronym/abbreviation identification: For example: “chf” for congestive heart failure
- Part-of-speech tagging

```

Processing 00000000.tx.1: stomach pain

Phrase: stomach pain
>>>> Phrase
stomach pain
<<<<< Phrase
>>>> Candidates
Meta Candidates (Total=9; Excluded=1; Pruned=0; Remaining=8)
 1000 C0221512:STOMACH PAIN (Stomach ache) [Sign or Symptom]
 1000 C1963242:Stomach pain (Stomach Pain Adverse Event) [Finding]
 861 C0030193:PAIN (Pain) [Sign or Symptom]
 861 C1962977:Pain NOS (Pain NOS Adverse Event) [Finding]
 694 C0038351:STOMACH (Stomach) [Body Part, Organ, or Organ Component]
 694 C1278920:Stomach (Entire stomach) [Body Part, Organ, or Organ Component]
 694 C1517454:STOMACH (Gastric Tissue) [Tissue]
 694 C3714551:Stomach (Stomach structure) [Body Part, Organ, or Organ Component]
 638 E C1704242:Gastric (Gastric (qualifier value)) [Qualitative Concept]
<<<<< Candidates
>>>> Mappings
Meta Mapping (1000):
 1000 C1963242:Stomach pain (Stomach Pain Adverse Event) [Finding]
Meta Mapping (1000):
 1000 C0221512:STOMACH PAIN (Stomach ache) [Sign or Symptom]
<<<<< Mappings

```

Figure 3.1: MetaMap concept mapping for “stomach pain”. The MetaMap maps “stomach pain” to the concept “stomach ache” and the Semantic Type “Sign or Symptom”.

- Lexical variant lookup of input words in the SPECIALIST lexicon
- Candidate Generation: For each term, the MetaMap generates a set of candidate concepts from the Metathesaurus that matches with the terms. These candidate concept mappings are evaluated based on a weighted scoring method that assigns a score (between 0 and 1000) to candidates based on how well they match with input text. The MetaMap orders candidates from higher to lower score. The higher the score, the higher is the probability that concepts relate to the phrase.
- Concept Mapping: In this step, candidates found in the previous step are combined and evaluated to produce a final result that best matches the phrase text

As shown in Figure 3.1, output of the MetaMap consists of three parts: 1) The phrase itself; 2) a list of the candidate concepts from the Metathesaurus. (In addition, the preferred name of each candidate is displayed in parentheses); 3) the mappings, combinations of candidates matching as

```

Processing 00000000.tx.1: water in brain

Phrase: water in brain
>>>> Phrase
water in brain
<<<<< Phrase
>>>>> Candidates
Meta Candidates (Total=6; Excluded=1; Pruned=0; Remaining=5)
  790 C0043047:WATER (Water) [Inorganic Chemical,Pharmacologic Substance]
  790 C0599638:WATER (Drinking Water) [Substance]
  790 C1550678:Water (Water Specimen) [Inorganic Chemical]
  718 E C0443350:Watery [Qualitative Concept]
  623 C0006104:BRAIN (Brain) [Body Part, Organ, or Organ Component]
  623 C1269537:Brain (Entire brain) [Body Part, Organ, or Organ Component]
<<<<< Candidates
>>>>> Mappings
Meta Mapping (746):
  790 C0599638:WATER (Drinking Water) [Substance]
  623 C0006104:BRAIN (Brain) [Body Part, Organ, or Organ Component]
Meta Mapping (746):
  790 C0599638:WATER (Drinking Water) [Substance]
  623 C1269537:Brain (Entire brain) [Body Part, Organ, or Organ Component]
Meta Mapping (746):
  790 C0043047:WATER (Water) [Inorganic Chemical,Pharmacologic Substance]
  623 C0006104:BRAIN (Brain) [Body Part, Organ, or Organ Component]
Meta Mapping (746):
  790 C0043047:WATER (Water) [Inorganic Chemical,Pharmacologic Substance]
  623 C1269537:Brain (Entire brain) [Body Part, Organ, or Organ Component]
Meta Mapping (746):
  790 C1550678:Water (Water Specimen) [Inorganic Chemical]
  623 C0006104:BRAIN (Brain) [Body Part, Organ, or Organ Component]
Meta Mapping (746):
  790 C1550678:Water (Water Specimen) [Inorganic Chemical]
  623 C1269537:Brain (Entire brain) [Body Part, Organ, or Organ Component]
<<<<< Mappings

```

Figure 3.2: MetaMap concept mapping for “water in brain”.

much of the phrase as possible.

MetaMap is highly configurable and provides various options for processing text and generating the mapping. For example, the output can include concept unique identifiers (CUI) and Semantic Types for the concepts. Also we can restrict concept mapping to certain vocabularies.

3.7.3 Concept Identification Challenge

While processing multi-word terms, sometimes the MetaMap does not map concepts properly. For example, the phrase “water in brain” is mapped to “water” (Drinking water) [substance] and “brain” (brain) [body part] (Figure 3.2)

In order to address such challenges, we 1) incorporated advanced text analytics techniques in the MetaMap processing and 2) used consumer health vocabulary (CHV) in the UMLS.

Advanced Text Analytics for Concept Identification

Following are some of the text analytics techniques that we utilized to improve the performance of the MetaMap in concept identification task.

- **Word Sense Disambiguation (WSD)**

Word sense disambiguation is a process of identifying the meaning of a term in context [Stevenson and Wilks 2003]. A word can have multiple interpretations based on the context in which it is used. For example, “discharge from hospital” versus “discharge from wound”. WSD is classic problem in the NLP research community. WSD is an important problem in the health domain as well as in UMLS. For example, the term “cold” can be interpreted 4 different ways in UMLS.

1. Cold (Cold Sensation) [Physiologic Function]
2. Cold (Cold Temperature) [Natural Phenomenon or Process]
3. Cold (Common Cold) [Disease or Syndrome]
4. Cold (Upper Respiratory Infections) [Disease or Syndrome]

To address the WSD problem, a range of approaches have been developed, including statistical ML techniques (supervised, semi-supervised, and unsupervised), linguistic techniques, and dictionary-knowledge based techniques. With the WSD module, the MetaMap generates mappings for the terms considering the surrounding text.

- **Term processing**

By default, MetaMap chunks its input into phrases, (noun phrases, prepositional phrases, etc.) each of which is analyzed separately. With the term processing module, MetaMap process each input record, as a single phrase, in order to identify more complex Metathesaurus terms.

- **Allowing concept gaps**

With this module, MetaMap can retrieve Metathesaurus candidates with gaps. For example, the text “obstructive apnea” will map to the concepts “obstructive sleep apnea” and “obstructive neonatal apnea”, which are considered too specific for normal processing.

There are other text analytics components that can be used for input processing which can “ignore word order”, support “overmatching of terms” and “composite phrase (identify concepts with multiple concepts)”.

Consumer health Vocabulary

Laypersons (“consumers”) often have difficulty finding, understanding, and acting on health information due to gaps in their medical domain knowledge. While health domain experts have foundational medical domain knowledge based on formal education and professional experience, laypersons have some socially and culturally derived notions of health and illness acquired from formal and informal sources (e.g., media exposure) and unique personal experiences [Zeng and Tse 2006]. Thus consumers use words and phrases (expressions) to describe health-related concepts that frequently differ from those used by professionals, such as “hair loss” for “alopecia”. In the current example, “water in brain” is actually a consumer-oriented term for a medical condition, “hydrocephalus”. Consumer health vocabularies link terms used by laymen to medical terms in the UMLS Metathesaurus. UMLS contains one CHV that maps consumer-oriented terms to UMLS Metathesaurus terms.

With these advanced text processing components and Consumer Health Vocabulary, MetaMap correctly identifies medical concept for “water in brain”. (Figure 3.3)

Next challenge is that CHV in UMLS is not comprehensive.

For example, for a search query “water on the knee” even with advanced text processing and CHV, MetaMap maps it to “Water thick-knee” (*Burhinus vermiculatus*) [Bird]. (Figure 3.4) This vocabulary gap is an even more serious problem for health search intent mining problem since a large

```

Processing 00000000.tx.1: water in brain

Phrase: water in brain
>>>> Phrase
water in brain
<<<<< Phrase
>>>> Candidates
Meta Candidates (Total=7; Excluded=1; Pruned=0; Remaining=6)
  790 WATER (Water) [Inorganic Chemical,Pharmacologic Substance]
  790 WATER (Drinking Water) [Substance]
  790 Water (Water Specimen) [Inorganic Chemical]
  718 E Watery [Qualitative Concept]
  623 BRAIN (Brain) [Body Part, Organ, or Organ Component]
  623 Brain (Entire brain) [Body Part, Organ, or Organ Component]
  580 Water on the brain (Hydrocephalus) [Disease or Syndrome]
<<<<< Candidates
>>>> Mappings
Meta Mapping (774):
  580 Water on the brain (Hydrocephalus) [Disease or Syndrome]
<<<<< Mappings

```

Figure 3.3: MetaMap correct concept mapping for “water in brain”.

```

Processing 00000000.tx.1: water on the knee

Phrase: water on the knee
>>>> Phrase
water on the knee
<<<<< Phrase
>>>> Candidates
Meta Candidates (Total=9; Excluded=1; Pruned=0; Remaining=8)
  770 WATER (Water) [Inorganic Chemical,Pharmacologic Substance]
  770 WATER (Drinking Water) [Substance]
  770 Water (Water Specimen) [Inorganic Chemical]
  699 E Watery [Qualitative Concept]
  604 Knee [Body Part, Organ, or Organ Component]
  604 Knee (Knee joint) [Body Space or Junction]
  604 Knee (Entire knee region) [Body Location or Region]
  604 Knee, NOS (Knee region structure) [Body Location or Region]
  591 Water thick-knee (Burhinus vermiculatus) [Bird]
<<<<< Candidates
>>>> Mappings
Meta Mapping (763):
  591 Water thick-knee (Burhinus vermiculatus) [Bird]
<<<<< Mappings

```

Figure 3.4: MetaMap correct concept mapping for “water on the knee”.

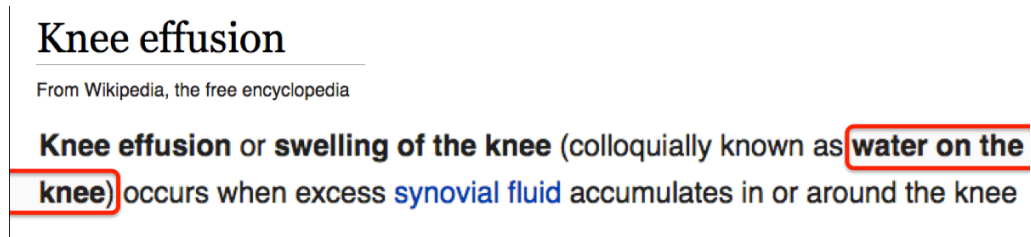


Figure 3.5: a snippet from a Wikipedia article on “knee effusion”.

of portion of health search queries are submitted by laymen. A key problem with a layperson’s query is that due to lack of knowledge about proper medical concepts (e.g. concepts from UMLS Metathesaurus) that can express his health information needs, the user may use common language to indirectly describe the concepts. To alleviate this problem, we leveraged crowd sourced knowledge from Wikipedia to improve the coverage of consumer health vocabularies. Wikipedia is the largest open access online encyclopedia. It is the one of the most-used online medical resources for both patients and healthcare professionals [Brokowski and Sheehan 2009].

Motivating Example

Here is a snippet from a Wikipedia article on “Knee effusion” (Figure 3.5). The article mentions alternate terms for “knee effusion” i.e. “swelling of the knee” and “water on the knee”. This knowledge helps us to map the consumer-oriented term “water on the knee” to the medical term “knee effusion”. Given “knee effusion”, the MetaMap correctly identifies it as “Disease or Syndrome” concept. Thus, Wikipedia can be a great knowledge source that we can leverage to improve the coverage of consumer health vocabularies and in turn to solve the health search intent mining problem.

3.8 Consumer Health Vocabulary Generation Using Wikipedia

The gap between lay and professional health terminologies has long been identified as one of the significant barriers to the empowerment of healthcare consumers. Studies suggest that lay people have difficulty understanding medical jargon [Chapman et al. 2003], and this affects their ability to search health-related information online, comprehend health information, and communicate effectively with their health providers. Consumer health vocabularies (CHVs) have been developed to [MacLean and Heer 2013]:

- Narrow the knowledge gaps between consumers and providers.
- Improve search and retrieval of health content.
- Improve comprehension of medical information from various Internet and printed sources for laymen.
- Aid consumer health informatics applications.
- Help consumer to communicate with health professional about their health conditions, treatment option and participate in decisions making process.

The medical informatics approach to solving the vocabulary problem involves building structured vocabularies of consumer health terms and mapping them to professional medical vocabularies [Zeng and Tse 2006]. The vocabulary development process typically involves building a structured vocabulary [Zeng and Tse 2006] by identifying consumer-oriented terms and “translating” them to terms used by health professionals by mapping consumer terms to their equivalents contained in professional medical controlled vocabularies (e.g., the UMLS Metathesaurus); for example, the layperson's nosebleed is a physician's epistaxis. Currently there are two open access consumer health vocabularies: the MedlinePlus Consumer Health Vocabulary, and the open and collaborative Consumer Health Vocabulary(OAC) CHV which was included in UMLS as of May 2011.

To date, most research in this area has focused on uncovering new terms to add to the (OAC) CHV. Researchers have used multiple data sources to identify consumer-oriented terms such as MedlinePlus search query logs [Zeng and Tse 2006] and patient defined data from PatientsLikeMe [Doing-Harris and Zeng-Treitler 2011]. These approaches generate a list of candidate terms, which are further, added to (OAC) CHV after manual review by health professionals. This approach is tedious and not scalable. Controlled vocabularies require maintenance and updating due to the continuing evolution of language itself [Hurford et al. 1998]. Consumer Health Vocabularies are no exception. As new findings emerge, new words are added to the vocabulary. In healthcare especially, there is a constant stream of new names (e.g., new medications, disorders, and tests) [Doing-Harris and Zeng-Treitler 2011]. Subsequently, CHV should also be kept updated with emerging health terms. To address these challenges, we leveraged crowdsourced knowledge from Wikipedia that is being continuously updated.

Wikipedia is the largest and the most visited online encyclopedia. It is widely regarded as a high quality, authoritative encyclopedia. It contains more than 5 million articles in English. One of the most compelling explanations for Wikipedia's success is, in short, "the wisdom of the crowds". Wikipedia is a very dynamic and fast growing resource (more than 20,000 new articles per month) articles about newsworthy events around the world are often added within a few days after their occurrence. Studies have found that its content is of comparable quality to traditional encyclopedias [Giles 2005], and that vandalism and inaccuracies are often reverted within a matter of minutes [Kittur et al. 2007; Arazy and Nov 2010]. Wikipedia includes nearly every aspect of human knowledge ranging from art and technology to health. Rich knowledge from Wikipedia has spurred development for variety of knowledge-bases such as YAGO [Suchanek et al. 2007] and DBpedia [Auer et al. 2007] and the knowledge-driven applications, Wolfram Alpha¹⁵, IBM Watson¹⁶, and Google knowledge graph¹⁷.

¹⁵<https://www.wolframalpha.com/>

¹⁶<http://www.ibm.com/smarterplanet/us/en/ibmwatson/>

¹⁷<https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>

- ▼ Diseases and disorders (38 C, 52 P)
 - ▼ Diseases and disorders by system (18 C, 1 P)
 - ▼ Blood disorders (9 C, 26 P)
 - ▶ Albumin disorders (4 P)
 - ▶ [Coagulopathies](#) (1 C, 55 P)
 - ▶ Deaths from blood disease (6 C, 14 P)
 - ▶ Haemorrhagic and haematological disorders (1 C, 1 P)
 - ▼ Hematologic neoplasms (2 C, 2 P)
 - ▼ Hematologic malignant neoplasms (3 C, 1 P)
 - ▼ Leukemia (5 C, 17 P)
 - ▼ Acute leukemia (2 C)
 - ▶ Acute lymphocytic leukemia (1 C, 1 P)
 - ▶ Acute myeloid leukemia (1 C, 1 P)

Figure 3.6: Wikipedia category hierarchy.

Wikipedia is organized hierarchically as an ontology. Each Wikipedia article is a single Web page and usually describes a single topic. Each Wikipedia article may be linked to other related articles by hyperlinks. The majority of Wikipedia pages have been manually assigned to one or more categories that represent the major topic of the article. These categories are organized and structured to allow users to browse their way around to find related information. They have a hierarchical structure (Figure 3.6). For example, Health and fitness –> Disease and disorders –> blood disorders –> leukemia. There are 12 parent categories on Wikipedia (Figure 3.7).

Going back to our CHV problem, community-generated text on Wikipedia could serve as a valuable resource to extract laypersons expressions of medical concepts (i.e., consumer terms) and their corresponding professional expressions. Wikipedia is the one of the most-used online medical

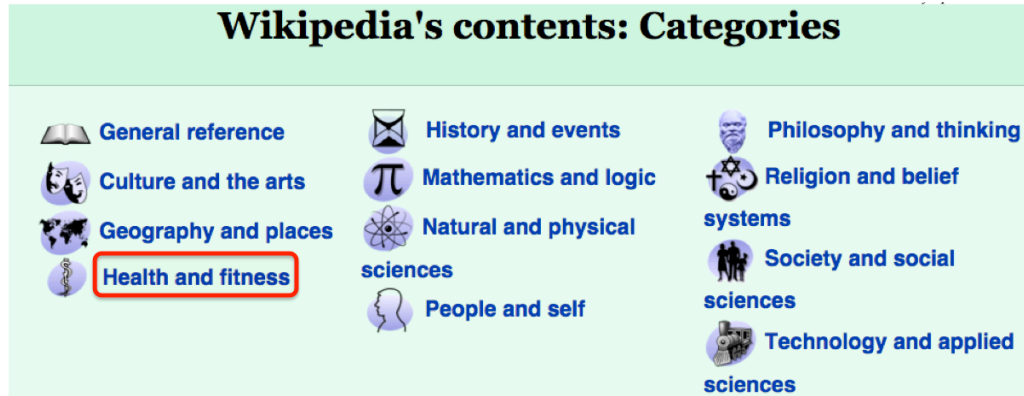


Figure 3.7: Parent categories on Wikipedia category hierarchy

resources, both for patients and healthcare professionals¹⁸. Wikipedia is freely accessible and often ranked in the top 10 results in Web search. Wikipedia provides complex health information in a simplified way, which makes it appealing for patients, caregivers and healthcare professionals. As shown in the motivating example (knee effusion, Figure 3.5), Wikipedia health articles tend to link consumer-oriented terms with health professionals' terminology using some semantic relationships (e.g., “Epistaxis, also known as a nosebleed”). Such knowledge makes Wikipedia very exciting resource for CHV generation. In this research, we exploited these relationships and this knowledge from Wikipedia to generate a consumer health vocabulary.

3.8.1 Approach

Let us look at two snippets from Wikipedia.

Snippet 1: Hair loss, also known as alopecia or baldness, refers to a loss of hair from the head or body.

Snippet 2: Knee effusion or swelling of the knee (colloquially known as water on the knee) occurs when excess synovial fluid accumulates in or around the knee joint.

¹⁸<http://m.nextgov.com/health/2014/02/wikipedia-massively-popular-yet-untested-doctor/79154/>.

Pairs	Term	Semantic Relation	Term
1	Hair loss	also known as	alopecia
2	Hair loss	also known as	baldness
3	Hair loss	refers to	loss of hair from the head/body
4	Knee effusion	colloquially known as	water on the knee
5	swelling of the knee	colloquially known as	water on the knee
6	Knee effusion	same as	swelling of the knee

Table 3.3: Candidate term pairs from Wikipedia snippets

In both the snippets, we can identify pairs of two terms that are related by semantic relationships as show in the Table 3.8.1

Although, using Wikipedia, we can generate semantically related candidate term pairs (e.g. {hair loss, alopecia}, {hair loss, baldness}, {knee effusion, water on the knee}, {swelling of the knee, water on the knee }), we can not identify CHV terms as Wikipedia does not state which term is consumer-oriented and which one is a medical professional term.

Thus, we can slice the problem of generating consumer health vocabulary using Wikipedia into the following two subproblems:

- To generate set of candidate pairs from health related Wikipedia articles.
- To identify consumer-oriented terms (henceforth referred to as CHV term) and health professional medical terms (henceforth referred to as medical term) from the set of candidate pairs.

3.8.1.1 Candidate Pair Generation from Health Related Wikipedia Articles

- **Identification of health-related Wikipedia articles**

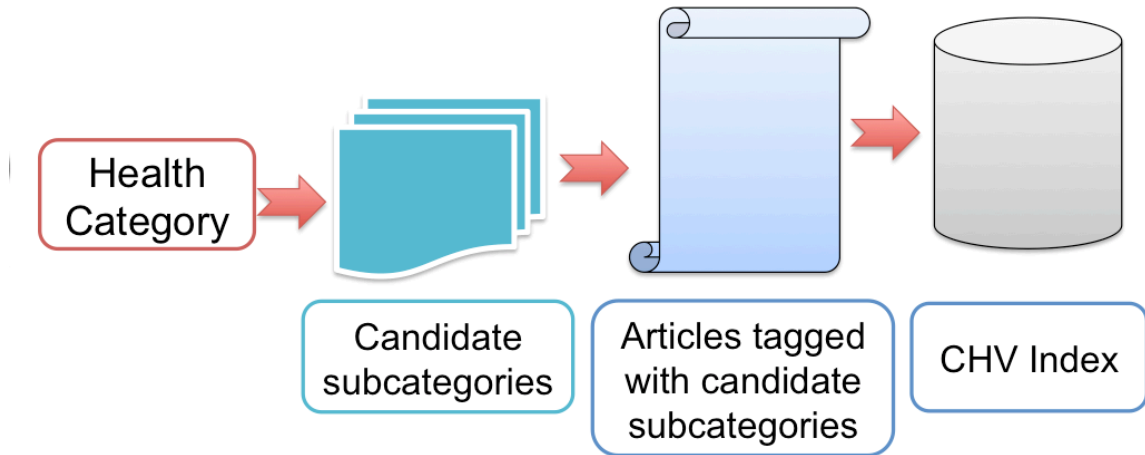


Figure 3.8: Approach for generating CHV

As mentioned in the previous section, Wikipedia content is organized in the form of Wikipedia articles. Each Wikipedia article usually describes a single topic and is manually labeled with one or more categories that represent the major topic(s) of the article. These categories are organized hierarchically with parent-child relationship. There are 12 parent categories on Wikipedia (Figure 3.7), and these categories further have sub-categories, which in turn have sub-categories.

For example:

Health – > Diseases and disorders – > Infectious diseases – > Bacterial diseases – > Animal bacterial diseases – > Cholera

Wikipedia data can be publically accessible and can be downloaded in XML format. The complete Wikipedia category hierarchy can be obtained by repeatedly traversing the sub-category links. We used an external Wikipedia tool, CatScan¹⁹, to collect a list of all sub-categories for the “Health” category down to a depth of three. CatScan searches an article category (and its subcategories) according to specified criteria to find articles, stubs, images, and categories²⁰.

Then we checked all the Wikipedia articles in the English language and selected the articles that

¹⁹<http://tools.wmflabs.org/catscan2/catscan2.php>

²⁰<https://meta.wikimedia.org/wiki/CatScan>

are tagged with, at least one of the candidate subcategories. There were total 1,593 candidate subcategories and 36K Wikipedia articles on health. We discarded articles that were not related to medical health, such as health by country, healthcare laws, health standards, and hospitals.

- **Extraction of candidate pairs**

All Wikipedia articles follow a consistent structure and format style. The articles that are not well formatted as per the Wikipedia guidelines get removed from Wikipedia. The first sentence of a Wikipedia article explains the topic of the article in simple terms. Also if the topic of the article has synonyms or alternate names or forms (e.g. spellings), then they also appear in boldface in the first sentence. Following are some guidelines from Wikipedia about formatting the first sentence

- Only the first occurrence of the title and significant alternative titles (which should usually also redirect to the article) are placed in bold²¹.
- If the subject of the page has a common abbreviation or more than one name, the abbreviation (in parentheses) and each additional name is given in boldface on its first appearance²².

For example:

Snippet 1: **Hair loss**, also known as **alopecia** or **baldness**, refers to a loss of hair from the head or body.

Snippet 2: **Knee effusion** or **swelling of the knee** (colloquially known as **water on the knee**) occurs when excess synovial fluid accumulates in or around the knee joint.

In the first snippet, “Hair loss” is the title of the Wikipedia article and “alopecia” and “baldness” are the synonyms or alternate names. Similarly, in the second snippet, “Knee effusion” is the title

²¹https://en.wikipedia.org/wiki/Wikipedia:Manual_of_style/Lead_section

²²https://en.wikipedia.org/wiki/Wikipedia:Manual_of_style/Lead_section#First_sentence

Wikipedia Patterns		
also called	commonly called	colloquially known as
also known as	commonly known as	sometimes called
also referred to as	commonly termed	sometimes known as
also termed	previously known as	sometimes termed
commonly referred to as	colloquially referred to as	sometimes referred to as

Figure 3.9: Wikipedia formatting patterns that are used to extract candidate pairs

of the Wikipedia article and “swelling of the knee” and “water on the knee” are the synonyms or alternate names. This consistent formatting style of Wikipedia articles can be leveraged for automated text mining and to generate candidate pairs. With the help of Wikipedia article formatting guidelines, we developed a pattern-based information extractor. The information extractor first splits each Wikipedia article on health into sentences (using OpenNLP) and then selects the first sentence to analyze further. Next, each sentence is processed by a pattern-based matcher algorithm that extracts candidate pairs from the sentences. In this step, the algorithm extracted total 9,030 candidate pairs from the Wikipedia articles.

3.8.1.2 Identification of CHV and medical terms from candidate pairs

Most of the previous approaches [Zeng and Tse 2006; Doing-Harris and Zeng-Treitler 2011] relied on manual review of candidate pairs to label them as CHV terms. The manual review task involves review of terms by health domain experts and with general agreement the terms are labelled as CHV terms. This approach is tedious and time consuming. Also, this approach is not scalable and feasible all the time. VGV [Vydiswaran et al. 2014], labeled a term as a CHV or medical term based on its probability of presence either in consumer-oriented texts (e.g. online health discussion forums) or

medical professional texts (e.g. PubMed articles). In this research, we mapped all the terms from candidate pairs to the UMLS Metathesaurus using the advanced text analytics of MetaMap.

Following are the three scenarios by which terms from candidate pairs are mapped to the UMLS Metathesaurus:

In scenario 1, both terms from the candidate pair are present in the UMLS Metathesaurus. For example, both terms from the {hair loss, alopecia} candidate pair are present in the UMLS Metathesaurus. In such cases, we do not need further processing and we discard the candidate pairs (total 5,418 pairs).

In scenario 2, both terms from the candidate pair are not present in the UMLS Metathesaurus. For example, both terms from the {hospital trust, acute trust} candidate pair are not present in the UMLS Metathesaurus. We hypothesize that since both terms are not part of UMLS they may be relevant to health topic in general but not that relevant to clinical health. In such cases, we do not need further processing and we discard candidate pairs (total 2784 pairs).

In scenario 3, one term from the candidate pair is present in the UMLS Metathesaurus and other term is not present (total 828 pairs). For example, from the {knee effusion, water on the knee} candidate pair, “knee effusion” is present in the UMLS Metathesaurus and “water on the knee” is not present in the UMLS Metathesaurus. We hypothesize that the term that present in the UMLS Metathesaurus is a medical term (e.g. knee effusion) and the term that is not present in the UMLS Metathesaurus is a CHV term (e.g. water on the knee). We empirically evaluated our hypothesis by querying terms from randomly selected candidate pairs on professional medical resources (e.g. PubMed) and consumer-oriented resources (e.g. forum, blogs).

We created an index of CHV terms and their medical terms. In the preprocessing step of the health search intent mining problem, we replace all the CHV terms from search queries with their medical terms. For example, a search query “symptoms for water on the knee” is replaced with “symptoms for knee effusion”. This crucial step helped us to improve health search intent mining approach.

3.9 The Corpus

For the experiments and evaluations, we selected health search queries related to chronic diseases.

3.9.1 Rationale for Data Selection

Chronic diseases, such as cardiovascular disease, stroke, cancer, chronic respiratory diseases and diabetes, are by far the leading cause of mortality in the world. Following are some facts about chronic diseases compiled from the Center for Disease Control and Prevention (CDC²³)

- As of 2012, about half of all adults in the United States, 117 million people, had one or more chronic diseases.
- One of four adults had two or more chronic diseases and the percentage of the US population living with chronic disease keeps increasing.
- In the United States, chronic diseases are the leading cause of death (7 in every 10 deaths) for both men and women.
- Two chronic diseases, cardiovascular disease and cancer, together accounted for nearly 48% of all deaths.
- The US spends 75% of healthcare dollars for the treatment of chronic diseases.

Chronic diseases are common across all socioeconomic groups and demographics, including all age groups, genders, and ethnicities. Most chronic diseases require lifelong care and the patient is in charge of managing the disease through self-care (such as diet, exercise and other healthy lifestyle choices). Prior studies [Ayers and Kronenfeld 2007; Fox and Duggan 2013] have shown that online resources are a significant information supplement for the patients with chronic conditions. As the percentage of people suffering from chronic diseases is very high, the number of people using the

²³<http://www.cdc.gov/chronicdisease/overview/index.htm>

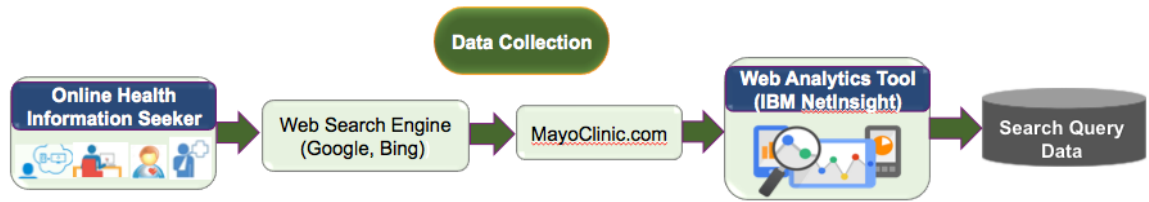


Figure 3.10: Search query data collection at Mayo Clinic

Internet to search and learn about them is also large. According to a 2013 Pew Survey [Fox and Duggan 2013], one in two American adults living with chronic diseases have gone online to find out information about a specific medical condition. Given the impact of chronic diseases on consumers' life and the significant search traffic for chronic diseases, this motivated us to select chronic diseases for this research. First, we conducted experiments on a cardiovascular diseases (CVD) dataset and then tested the approach on cancer and diabetes datasets.

3.9.2 Data Source

We have collected CVD-related search queries originating from Web search engines that direct online health information seekers to Mayo Clinic's consumer health information portal (MayoClinic.com), which is one of the top online health information portals within the United States. The MayoClinic.com portal provides up-to-date, high-quality online health information produced by professional writers and editors. Mayo Clinic's 2014 Web analytics statistics indicate that the MayoClinic.com portal is on average visited by millions of unique visitors every day and around 90% of the incoming traffic originates from search engines. This significant traffic to the portal provides us with an excellent platform to conduct our study.

3.9.3 Dataset Creation

The MayoClinic.com Web Analytics tool (IBM Netinsight on Demand) keeps detailed information about Web traffic such as input search query, time of visit and landing page. MayoClinic.com has

several CVD-related webpages that are organized by health topics and disease types. Using the Web Analytics tool, we obtained 10 million CVD-related anonymized search queries originating from Web search engines that “land on” CVD webpages within MayoClinic.com and are related to CVD. These queries are in English and were collected between September 2011-August 2013. Our final analysis dataset consists of 10,408,921 CVD related search queries, which is a significantly large dataset for a single class of diseases.

3.9.4 Gold Standard Dataset Creation

We randomly selected 2,000 search queries from the analysis dataset. Two domain experts manually annotated 2,000 search queries by labeling one search query with zero, one, or more than one intent classes. The annotators first discussed and agreed upon the annotation scheme. To reduce the probability of human errors and subjectivity, the two annotators discussed and annotated each query and created a gold standard dataset with 2,000 search queries, which was further divided into training and testing datasets with 1,000 search queries each. Training dataset was used to develop rule-based classification approach.

3.10 Data Preprocessing

In data preprocessing, first we performed data cleaning (e.g. removed all non-English search queries) and stop word removal. Then we corrected the misspellings, replaced CHV terms from the search queries with their medical terms and finally annotated the search queries with MetaMap.

Misspelling Correction

Online health information seekers occasionally make spelling mistakes while searching for health information. If a search query contains spelling mistakes then the annotation tool, MetaMap, may not map misspelled terms from the query to the UMLS Metathesaurus concepts. In order to correct

such errors, we used a dictionary-based approach. We first generated a dictionary of words using the Zyzyyva wordlist²⁴, the Hunspell dictionary²⁵, and its medical version (OpenMedSpell²⁶), comprising a total of 275,270 unique words. We used this dictionary with a spell corrector algorithm to correct the misspellings in the CVD search queries.

Replace CHV terms with medical terms

As discussed in Section 3.8, some CHV terms are not mapped in the UMLS Metathesaurus concepts. Since a large portion of health queries are submitted by non-experts, the prevalence of CHV terms in search queries also tends to be high. Thus, to alleviate this problem, we leveraged crowd-sourced knowledge from Wikipedia and created an index of CHV terms and their medical terms (Section 3.9). In this step, we replaced all the CHV terms from search queries with their medical terms from the UMLS Metathesaurus.

Data Annotation with the UMLS Metathesaurus Concepts

We utilized UMLS MetaMap tool for annotating the search queries with UMLS concepts and Semantics Types. We can access MetaMap by installing the MetaMap server. Once the server is running, it can be queried with text input and the server returns the UMLS concepts, their Semantic Types, Concept Unique Identifiers (CUIs), and other details for the terms in the text.

MetaMap Usage Challenge And Solution

Although MetaMap is a great tool for annotating medical concepts from the search queries, it is very inefficient in terms of processing. For example, just to annotate the 100,000 search queries using a single node MetaMap server, it takes couple of hours. Since the size of our dataset was fairly large (10 million), it was estimated that MetaMap would take a significant amount of time

²⁴<http://www.zyzyyva.net/wordlists.shtml>

²⁵<http://hunspell.sourceforge.net/>

²⁶http://www.e-medtools.com/Hunspell_openmedspell.html

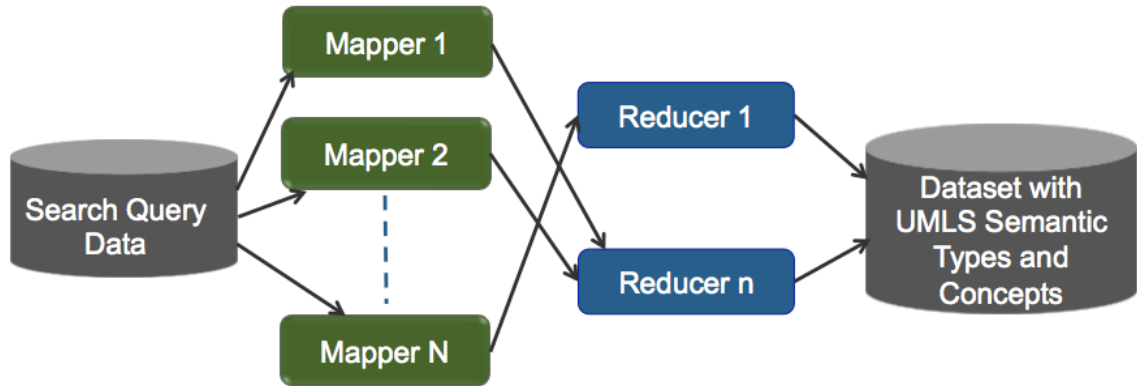


Figure 3.11: Hadoop-MapReduce framework with 16 nodes for MetaMap implementation

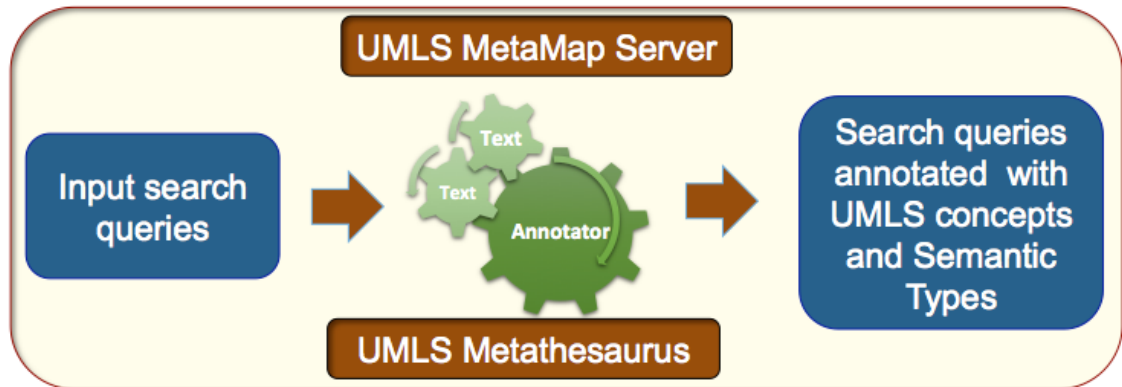


Figure 3.12: Functional overview of a Mapper

(in days) to annotate 10 million search queries. To address this challenge and to improve data annotation speedup, we implemented a 16 node Hadoop-MapReduce framework with a MetaMap server installation on each node (Figure 3.11).

Once the search query data Q is submitted to the Hadoop framework, it is divided into $dc = Q/16$ chunks, one for each node. Each node further splits its data chunk into dc/N , where, N is number of mapper on a node. Each node has multiple mappers (N), which process the search queries using a MetaMap server installed on the node (Figure 3.11). Mappers output the search query with UMLS Concepts and Semantic Types and reducers consolidated mappers' output. With this framework,

we observed a very significant improvement in the data processing time.

3.11 Classification Approach

To classify 10 million search queries into zero or more intent classes from a set of 14 intent classes, we developed a semantics-based classification approach. In the data preprocessing step, we annotated all the search queries with the UMLS Metathesaurus concepts (henceforth referred as Semantic Concepts) and Semantic Types. In the classification approach, we iteratively developed classification rules using the labeled search queries from the training dataset. Each rule is evaluated separately.

3.11.1 Classification Rules

- **Semantic Types (ST)**

As mentioned earlier, STs are broad subject categories and each concept in the Metathesaurus is assigned one or more STs. There are total 135 STs in UMLS. For medical text (e.g., search queries, EHR documents) classification prior approaches have used only Semantic Types [Natarajan et al. 2010; Denecke and Nejd1 2009; Humphrey et al. 2006; Pratt and Fagan 2000; Pratt and Wasserman 2000]. In our experiments, we used STs as a baseline approach to classify the search queries. Some of the STs can be directly mapped to our intent classes. Based on the description of intent classes and STs, we assigned semantically relevant STs to matching intent classes. For example, the following three semantic types; DIAP - Diagnostic Procedure, LBPR - Laboratory Procedure, and LBTR - Laboratory Test Result are semantically very relevant to ‘Test and Diagnosis’ intent class. As a baseline approach, first we classified search queries based on STs. A classification approach based only on STs had 54.32% precision, 62.03% recall and a 57.91% F1 score.

- **ST + Semantic Concepts (SC)**

To improve the baseline approach, we assigned semantically relevant, generic, and frequent SCs

to related intent classes. For example, ‘medication’, ‘medicine’, ‘drugs’, ‘dose’, ‘dosage’, ‘tablet’, ‘pill’ are SCs that are semantically relevant to the ‘Drugs and Medication’ intent class. For a few intent classes (e.g., ‘Food and Diet’), there are certain concepts that are closely associated with the intent class and yet are not mapped to the selected semantic type. For example, ‘FOOD’ ST does not include concepts such as ‘meal’, ‘menu’, ‘diet’, ‘recipe’ and ‘lunch’. Note that, although these concepts are related to ‘FOOD’ ST they are not actually food items. Thus, in UMLS they are not labelled with FOOD ST. Whereas, in the context of search query intent mining, search queries with concepts are related to the ‘FOOD and Diet’ intent class. A classification approach based on STs and SCs improved the baseline performance and had 65.34% precision, 68.22% recall, and a 66.74% F1 score.

- **ST + SC + Keywords (KW)**

We noticed that certain terms appear very frequently in search queries but are not part of concepts in the UMLS Metathesaurus. For example, ‘living with’ (living with heart attack, living with diabetes). Thus, to further improve the performance of the classification approach, we considered certain keywords associated with the intent classes. With keywords, the performance of the classification approach improved marginally and had 67.22% precision, 69.23% recall, and a 68.21% F1 score.

- **ST + SC + KW - ST and ST + SC + KW - ST - SC**

After analyzing the classification of the search queries at this step, we observed that the exclusion of certain STs and SCs from intent classes can be helpful. A few STs include some undesired concepts (in the context of our customized classification, not in terms of the UMLS concept hierarchy). For example, STs ‘ORCH - Organic Chemical’, ‘CLND - Clinical Drug’ and ‘PHSU - Pharmacologic Substance’ are associated with the ‘Drugs and Medication’ intent class (Figure 3.13). These STs include some concepts that are not considered as drugs by a consumer/lay population, such as caffeine, fruit, prevent, and alcohol. With the exclusion of STs

Intent Class	Classification Rule	Examples
Drugs and Medications	<ul style="list-style-type: none"> • $\{ST \cup SC \cup KW\} \setminus SC^*$ • ST: ORCH PHSU, CLND, PHSU • SC: medication, medicine, drugs, dose, dosage, tablet, pill • KW: meds • (Without) SC*: alcohol, caffeine, fruit, prevent 	<ul style="list-style-type: none"> • Medications for pulmonary hypertension • ibuprofen heart rate • Dextromethorphan blood pressure

Figure 3.13: Classification rule for Drugs and Medications intent class

the classification approach had 71.83% precision, 72.44% recall, and a 72.13% F1 score. Further, the classification performance improved with the exclusion of SCs and had 76.01% precision, 79.30% recall, and a 77.62% F1 score.

- **ST + SC + KW - ST - SC + Advanced Text Analytics (AdvTA)**

As mentioned in Section 3.7, we can improve the performance of MetaMap in the concept identification task by incorporating advanced text analytics modules such as word sense disambiguation, term pre-processing, allowing concept gaps, ignoring word order, and overmatching of terms and composite phrases (identifying concepts with multiple concepts). With advanced text analytics, the performance of the classification approach increased significantly and had 85.39% precision, 83.82% recall, and a 84.59% F1 score.

- **ST + SC + KW - ST - SC + AdvTA + CHV**

Some CHV terms are not mapped in the UMLS Metathesaurus. Using Wikipedia, we created an index of CHV terms and associated medical terms. The generated index was used in the preprocessing step to replace CHV terms from search queries with their medical terms. After using this CHV component, the performance of the classification approach increased and had

88.42% precision, 86.07% recall, and an 87.23% F1 score.

3.11.2 Classification Algorithm

Notations:

Let $\mathbf{Q} = \{q_0, q_1, \dots, q_i\}$ be the set of search queries

$\mathbf{ST} = \{t_0, t_1, \dots, t_u\}$ be the set of Semantic Types

$\mathbf{SC} = \{c_0, c_1, \dots, c_v\}$ be the set of semantic concepts

$\mathbf{IC} = \{ic_1, ic_2, \dots, ic_k\}$ be the set of intent classes for query Q

$\mathbf{P} = \{p_0, p_1, \dots, p_j\}$ be the set of query phrases extracted from Q

$\beta : \mathbf{Q} \rightarrow \mathbf{P}$, be the function that maps one query to a set of phrases

$\beta(q_i) = \{p_0, p_1, \dots, p_j\}, (j \geq 0)$

$\lambda : \mathbf{P} \rightarrow \mathbf{ST} \cup \mathbf{SC}$, be a function to assign a set of Semantic Types or concepts to a

query phrase $\lambda(p_j) = t_0, t_1, \dots, t_m, c_0, c_1, \dots, c_n$ with $t \in \mathbf{ST}, c \in \mathbf{SC}$

$\alpha(q_i) = U\lambda_j$ where $\lambda_j = \lambda(p_j)$ and $\beta(q_i) = p_0, p_1, \dots, p_j$, be the annotation that assigns a set of Semantic Types and concepts to a query

$\mathbf{R}(ic_k)$ = Rule function which returns the set of Semantic Types and concepts to be included for intent class ic_k

$\mathbf{R}'(ic_k)$ = Rule function which returns the set of Semantic Types and concepts to be excluded for intent class ic_k

3.12 Evaluations and Results

3.12.1 Classification Approach Evaluation

The classification rules developed in the previous section are evaluated on 1,000 search queries from the gold standard (testing) dataset. We used Macro Average Precision Recall as evaluation metric.

Macro Average Precision Recall is calculated by computing the average of precision and recall for

Algorithm 1 Intent classification algorithm

```

1:
2: for Query  $q_i \in Q$  do
3:    $ic_{q_i} = \{\}$  ▷ Intent class initialization for query  $q_i$ 
4:   for phrase  $p_j \in q_i$  do
5:     for intent class  $ic_k \in IC$  do
6:       if CLASSIFY( $p_j, ic_k$ ) then
7:          $ic_{q_i} \leftarrow ic_{q_i} \cup ic_k$ 
8:       end if
9:     end for
10:  end for
11: end for

```

Algorithm 2 Classification function

```

function CLASSIFY( $p_j, c_k$ )
2:  if  $(\lambda(p_j) \cap R(ic_k) \neq \emptyset \vee p_j \in R(ic_k)) \wedge (\lambda(p_j) \cap R'(ic_k) = \emptyset \wedge p_j \notin R'(ic_k))$  then
    $ic_{q_i} \leftarrow ic_{q_i} \cup ic_k$ 
3:  end if
end function

```

Rules	Precision	Recall	F1 Score
ST	0.5432	0.6203	0.5791
ST+SC	0.6534	0.6822	0.6674
ST+SC+KW	0.6722	0.6923	0.6821
ST+SC+KW-ST'	0.7383	0.7344	0.7363
ST+SC+KW-ST'-SC'	0.7601	0.793	0.7762
ST+SC+KW-ST'-SC'+AdvTA	0.8539	0.8382	0.8459
ST+SC+KW-ST'-SC'+AdvTA+CHV	0.8842	0.8607	0.8723

Table 3.4: Evaluation of the classification approach

Where, ST = Semantic Type,

SC = Semantic (UMLS) concepts,

KW = keyword,

AdvTM = Advanced Text Analytics, and

CHV = Consumer Health Vocabulary

each individual class. Macro averaging gives equal weight to each class. The Macro-average F1-Score is the harmonic mean of Macro Average Precision and Macro Average Recall. Based on the evaluation, our classification approach had very good Precision: 0.8842, Recall: 0.8642, and F-Score: 0.8723.

3.12.2 Classification Evaluation by Intent Classes

We also performed a precision and recall analysis for each intent class independently (Table 3.12.2) to check the performance of the classification approach for individual intent classes. The classification approach performs very well for most of the intent classes. We observed that one reason which affected the classification performance was the ambiguous interpretation of some of the concepts that sometimes may not be contextually correct-e.g. for the search query “nuts good for your

No.	Intent Classes	Precision	Recall	F-Score
1	Symptoms	0.9274	0.8042	0.8614
2	Causes	0.8861	0.9859	0.9333
3	Risks and Complications	1	1	1
4	Drugs and Medications	0.8582	0.935	0.895
5	Treatments	0.7083	0.9444	0.8095
6	Tests and Diagnosis	0.6389	1	0.7797
7	Food and Diet	0.9391	0.9558	0.9474
8	Living with	0.8659	0.9342	0.8988
9	Prevention	0.8333	1	0.9091
10	Side effects	1	1	1
11	Medical devices	0.8077	0.75	0.7778
12	Diseases	0.9291	0.7751	0.8451
13	Age-group References	1	0.8889	0.9412
14	Vital signs	0.8872	0.8669	0.8769
Macro Average Precision (0.8842),				
Recall (0.8607) and F-Score (0.8723)				

Table 3.5: Performance of the classification approach with respect to individual intent classes

heart”, MetaMap annotated “nuts” as FOOD as well as Medical Device (Nut - Medical Device Component or Accessory).

Using the classification approach, we classified 10,408,921 CVD search queries into 14 intent classes. Since a query can be classified into multiple classes (multi-label classification), the total number of queries in the Table 3.12.2 is 14.7 million. Based on Table 3.12.2, the most popular intent classes while searching for CVD information are ‘Diseases and Conditions’ and ‘Vital signs’.

One in every two searches is related to either ‘Diseases and Conditions’ or ‘Vital signs’. Due to close association of vital signs (such as blood pressure and heart rate) with CVD, online health information seekers might be searching for it frequently. Other popular intent classes that users search for include ‘Symptoms’, ‘Living with’, ‘Treatments’, ‘Food and Diet’, and ‘Causes’. Mostly, due to the chronic nature of the CVD and as the patients are in charge of managing the disease with day-to-day care, many CVD patients are searching for ‘Living with’ related search queries. As diet has a significant impact on the CVD, we observed large search traffic for the ‘Food and Diet’ category. Many consumers are also interested in learning about CVD ‘Treatments’, ‘Medical Devices’ (e.g. pacemaker), ‘Drugs and Medication’, and ‘Cause’. Although CVD can be prevented with some lifestyle and diet changes, interestingly, very few consumers search for CVD ‘Prevention’.

3.12.3 Distribution of Search Queries by Number of Classified Intent Classes

A search query can be classified into zero or more intent classes. Using our classification approach, we classified 92% of the 10 million CVD related queries into at least one intent class (Table 3.12.3). Most of the queries (around 88%) are classified into either one or two intent classes. Very few CVD queries (4.28%) are classified into 3 or more intent classes. Our approach did not classify 8.13% of the queries into any intent classes. After studying the unclassified search queries, we found that there are a few queries that do not fit into any of the selected 14 intent classes, such as cardiac surgeon, cardiology mayo, video on cardiovascular, pediatric cardiology, and orthostatic.

3.12.4 Evaluation with respect to three chronic diseases

One important constraint that we had for this multi-label classification approach was that the approach should classify health related search queries in a disease-agnostic manner. To evaluate the performance of the classification approach, we selected two other chronic diseases, diabetes, and cancer. For dataset creation and evaluations, we followed the same approach as described for

No	Intent Classes	Total Queries	Percentage Distribution
1	Diseases	4,232,398	40.66
2	Vital signs	3,455,809	33.2
3	Symptoms	1,422,826	13.67
4	Living with	1,178,756	11.32
5	Treatments	955,701	9.18
6	Food and Diet	779,949	7.49
7	Med Devices	665,484	6.39
8	Drugs and Medications	603,905	5.8
9	Causes	599,895	5.76
10	Tests & Diagnosis	344,747	3.31
11	Risks and Complication	277,294	2.66
12	Prevention	136,428	1.31
13	Age-group References	87,929	0.84
14	Side effects	25,655	0.25
	Total	14,766,776	141.87

Table 3.6: Classification of search queries by intent classes

Number of health Categories	Number of search queries	Percentage Distribution
0	845,744	8.13%
1	4,967,337	47.72%
2	4,149,803	39.87%
3	420,622	4.04%
4 and 5	25,415	0.24%
Total	10,408,921	100.00%

Table 3.7: Classification of search queries by intent classes

Dataset	Precision	Recall	F1-Score
Cardiovascular Diseases	0.8842	0.8642	0.8723
Diabetes	0.9274	0.8964	0.9116
Cancer	0.8294	0.7635	0.795

Table 3.8: Performance of the classification approach with respect to three major chronic diseases cardiovascular diseases. As shown in the table 3.12.4, the classification approach performs very well with respect to all three chronic diseases (CVD, diabetes and cancer). In fact, the performance of the classification approach for diabetes is even better than that for CVD. One major reason for the improved performance of diabetes is that the diabetes have fewer classes (like type 1 and type 2) and less ambiguity. While for CVD and cancer, there are many sub-classes (of diseases) and more ambiguities, such as ‘cancer’ – > disease, zodiac sign, ‘heart attack’ – > disease, song. In such cases, contextual cues from the surrounding text play a crucial role in word sense disambiguation. However due to limited length of the search queries, sometimes there are limitations on WSD. This evaluation of the classification approach for three major chronic diseases confirms that the classification approach can work reasonably well in a disease agnostic-manner.

3.13 Conclusion

In summary, the following are our major contributions in this work:

- We developed an approach to automatically identify health search intents from large-scale search logs in a disease-agnostic manner [Jadhav et al. 2014; Jadhav et al. 2014a; Jadhav et al. 2014; Jadhav et al. 2014b].
- We constructed a consumer health vocabulary that maps laymen terms to medical terms used by health professionals by parsing health related Wikipedia articles.
- In the MetaMap data processing, we used advanced text analytics techniques like word sense disambiguation and term processing, and utilized consumer health vocabulary to improve concept identification from the search queries.
- We developed a scalable MetaMap implementation using Hadoop-MapReduce framework to improve MetaMap's data annotation speedup.

4

A Hybrid Approach for Identification of Informative Tweets and Social Health Signals System

In this chapter, we will apply the search intent mining algorithm presented in the Chapter 3 on health related Twitter data. Since Twitter data is very noisy, we first addressed the problem of identification of informative tweets from noisy Twitter data. We used a hybrid approach consisting of supervised machine learning, rule-based classifiers, and biomedical domain knowledge to facilitate the retrieval of informative health information shared on Twitter in real time using Social Health Signals system.

4.1 Introduction

While users often turn to search engines to learn about health conditions, a surprising amount of health information is also shared and consumed via social media. Information behavior researchers have described two primary approaches for information acquisition [Lu 2012]. The first is intentional information acquisition, which involves the active seeking for information and is generally triggered by users' information need, e.g. information seeking using Web search. However, in many circumstances, users discover information on the Social Web merely by chance (i.e., accidental discovery of information [Lu 2012]). For example, a user may unexpectedly obtain certain information about a new clinical trial for diabetes patients while routinely checking his Twitter feeds. This experience of accidental information discovery refers to bumping into information (useful or personal interest-related) as opposed to intentionally looking for it [Erdelez 1997]. Social networking websites such as Facebook and Twitter provide excellent opportunities for accidental information acquisition. Through such websites, users may come across a great deal of unexpected useful information which can play an important role in their everyday information acquisition.

In the past few years, social media and especially the popular micro-blogging website Twitter have emerged as some of the major information sources that Web users are employing to keep up with the newest health information. A survey [Fox and Jones 2009] indicated that as many as 39% of online health information seekers have used social media, and a fraction of them had also followed their contacts' health experiences, posted their own health-related comments, gathered health information, or joined a health-related group. Other research has shown that people prefer search engines while seeking information for various sets of medical conditions, and prefer Twitter for sharing and learning about new health information [De Choudhury et al. 2014]. In some cases, people prefer Twitter as an information source, as compared to the traditional information sources (e.g. newspapers), since they can find timely information aggregated in one place information which they would not think to check for on the Internet of their own accord.

In many cases, the phenomenon of accidental information discovery is facilitated by the users' prior actions. For example, a person who is interested in keeping track of online health information may follow health-related Twitter accounts that can provide him the newest, reliable health information. This is also known as serendipity [Roberts 1989]; the chance of bumping into unexpected information can be increased by frequently interacting with other people or being exposed to an information-rich environment [Erdelez and Rioux 2000] (here health-related Twitter accounts). Currently, on Twitter, there are thousands of health-centric accounts that are followed by millions of users to keep up with health information. For example, some of the health-centric twitter accounts have more than a million of followers, such as @DrOz 3.81M, @goodhealth 3M, @WomensHealthMag 3.92M, @MensHealthMag 3.18M, @DailyHealthTips 2.9M, @MayoClinic 1.26M, and @WebMD 1.3M. These millions of followers indicate peoples interest in using the Twitter platform to keep track of health information.

However, the sheer volume of tweets on health topics is overwhelming. Hence, it is difficult to distill the most relevant tweets from the deluge of tweets while also filtering out tweets. Most of these tweets are highly personal and contextual. Therefore, most of them are neither interesting nor indeed meaningful to anybody outside of the author's circle of followers. According to [Naaman et al. 2010], less than 12% of the tweets are informative. Here, we define an informative health tweet as a tweet which conveys or points to useful health information of general interest (i.e., that is informative, useful, or beneficial to a general audience). There is no easy way to find informative tweets. Keyword-based search on Twitter does not consider the semantics of the query and returns all the chronologically ordered tweets containing the keyword; it does not rank tweets by considering the informativeness or reliability of the information. Here, relevancy is not an issue as all the tweets on a topic are relevant, but understanding the usefulness or informativeness of tweet is a problem. In most of the cases, a user has to go through all the tweets manually and has to depend on his/her own intellect, knowledge, and analytical capabilities to identify informative tweets. Since only a limited numbers of tweets are informative, finding informative tweets in a sea of millions of irrelevant chatter

remains a challenging research issue.

Furthermore, the informativeness of a tweet is very subjective. Twitter users produce diverse content ranging from news and events to conversations and personal status updates. Consider the following three examples scenarios:

1. If John tweeted: “I voted today!”, this tweet may be informative to John's school friend, Jenny, who is aware of the fact that there is an election in the school for a student representative role. However, the majority of the people will find this tweet uninformative (*personal context*).
2. Rob reads and shares an article on Twitter about the release of new medicine for patients with type 2 diabetes . People who do not have diabetes find this information irrelevant (*lack of interest in the topic*). While most of the people who have diabetes find this tweet informative, some people who are already aware of this information do not find the article informative (*novelty in the information or prior knowledge about the information*).
3. Also, sometimes “who has said it” can matter as much as or even more than “what has been said”. Advice about how to prevent diabetes from a diabetes specialist is more informative than advice from an ordinary person (*author's expertise or credibility in the information*).

To facilitate identification of informative and trustworthy content in tweets, it is crucial to develop an effective classification system that can objectively classify tweets as informative or uninformative . Thus, in order to address this problem, we have abstracted out the subjective nature of the informativeness problem and objectively studied what features contribute to informativeness. We developed a hybrid approach consisting of supervised machine learning and rule-based classifiers for the classification of informative vs. tweets. We leveraged biomedical domain knowledge to facilitate the retrieval of relevant and reliable health information shared on Twitter in real-time using a system called “Social Health Signals”. Moreover, we extended our search intent mining algorithm to classify health-related tweets into health categories that facilitate the browsing of the informative tweets and health news by health categories.

4.2 Approach

4.2.1 Data Collection

In this study, we have used messages shared on Twitter, i.e. tweets, as the data source. Twitter has 1.3 billion users, out of which 320 million are active users ¹. Each tweet fits within Twitter's 140-character limit and optionally contains URLs (links) as a pointer to an external piece of detailed information. Twitter users often use URL shortening services to make URLs shorter. Twitter offers a set of streaming APIs: public streaming, user streaming, and site streaming. We have used the public Twitter streaming API² to collect health-related tweets. The streaming API provides access to the public tweets based on the keywords. We selected tweets related to diabetes for the experiment and evaluation. One of the key reasons for selecting diabetes-related tweets for the experiments is to minimize noise in the dataset. Diabetes is one of the major chronic diseases. It has fewer subtypes and less ambiguous terms as compared to the other popular chronic diseases like cardiovascular diseases and cancer. We collected keywords related to diabetes by studying UMLS, diabetes forums, and diabetes-related tweets. Using the Twitter streaming API and diabetes-related keywords, we collected over 690,283 tweets over a period of 5 months.

4.2.2 Rule-based Filtering

First, we randomly selected 40K tweets from the dataset for the experiments to determine informative tweets. For each tweet, we identified its language and filtered out non-English tweets. This step reduced the experiment dataset from 40K to 29K tweets. Out of the 29K English tweets, 17.4K tweets contained at least one URL and 12K tweets did not contain any URL. After an empirical evaluation of the tweets without URLs, we noticed that most of the tweets were personal messages, jokes, and contained contextual information. Also, the reliability of the information mentioned in the

¹<http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/>

²<https://dev.twitter.com/streaming/>

tweets could not be verified as these tweets do not provide a reference for the information mentioned in the tweets (in the form of URL). Prior works in the field of tweet search have identified that the presence of a URLs in the tweets is the most effective feature of informative tweets [Duan et al. 2010; Massoudi et al. 2011]. Therefore, we only selected tweets with URLs (17.4K) for further study and then we performed the following filtering operations.

Duplicate tweet removal:

One of the features of Twitter is its retweet functionality that allows members to forward a tweet through their network. People often retweet to share an interesting piece of information on their network, to show agreement with the tweet content, and for a variety of other reasons [Boyd et al. 2010]. The practice of retweeting generates a significant number of duplicate tweets. In this step, we filtered out duplicate tweets and reduced the dataset from 17.4K tweets to 13.5K tweets.

Minimum length:

Due to Twitter's restriction, a tweet itself is limited to 140 characters. Tweets that only contain URLs but do not provide a brief summary about the topic of the URLs tend to be unreliable or spam. We should note that there might be false positives with this approach, but in the study we are focusing only on precision. We selected tweets that have at least 80 characters and 5 words apart from the URL. (The dataset size is 10.9K tweets).

Spelling mistakes:

Given the informal nature of Twitter and the high prevalence of slang and abbreviations, the percentage of tweets with spelling mistakes is high. At the same time, a well written and reliable information sharing tweet contains very few spelling mistakes. Considering the nature of Twitter text, we selected tweets that have at the most 2 spelling mistakes. Note that, on Twitter, people use the hashtag functionality to self label the topics of the tweets. A hashtag is a single (or a composite)

word, preceded by the # character e.g. #diabetes, #diabetesType1. We did not consider hashtags and URLs for the spelling mistakes. (The dataset size is 10.1K tweets.)

URL filtering:

Most of the URLs on Twitter are shortened using services such as Bitly³, TinyURL⁴, and Google URL shortener⁵. Since people use different URL shortening services, unless we expand the URLs, we cannot identify duplicates. We used an external library to expand shortened URLs. We filtered out broken and non-functional URLs. Finally, we filtered out duplicate URLs by retaining URLs from the tweets that have the maximum amount of words and minimum amount of spelling mistakes. (The dataset size is 8.2K tweets after filtering.)

PageRank:

The Google PageRank for a URL ranges from 0 to 10, with higher values indicating greater importance. The Google search algorithm uses more than 200 signals to calculate a website's PageRank, which indicates its overall importance, authority, and reliability. We utilized URL's PageRank as a reliability feature and retained the URLs that had a minimum PageRank of 5. (The dataset size is 6.3K tweets.)

Using the rule-based filtering (Table 4.2.2) , we reduced the experiment dataset from 40K tweets to 6.3K tweets (84.25% reduction in the dataset).

4.2.3 Classification

We used a supervised learning approach to classify tweets into informative and noninformative classes.

³<https://bitly.com/>

⁴<http://tinyurl.com/>

⁵<https://goo.gl/>

Filter	Description	Dataset size (in number of tweets)
Language	English	29,034
Tweet with URL		17,422
Duplicate tweet removal		13,573
Minimum length	- Number of words = 5 - Number of characters = 80	10,927
Spelling mistakes	Maximum 2	10,176
URL filtering	- Remove broken and not working URLs - Duplicate URLs	8,273
PageRank	Minimum 5	6,374

Table 4.1: Rule-based filtering

Gold Standard Dataset Creation

We randomly selected 3,000 tweets from the experimental dataset of 6.3K tweets. Since it is difficult to judge whether a tweet is informative, instead of using binary labels, we adapted scores from 1 (least informative) to 4 (extremely informative) for the labeling task. Three annotators first annotated 100 tweets together to agree on an annotation scheme. After the annotation scheme was finalized, the annotators independently assigned informativeness scores from 1 to 4 to the remaining tweets based on the tweet's content and URL. We then transformed the informative scores to binary labels (scores of 1 and 2 - non-informative; scores of 3 and 4 - informative). We selected the binary labels which were agreed upon by at least two annotators. In the labeled dataset, 33.6% were informative tweets and 66.4% were uninformative tweets. We divided the gold standard dataset into a training and testing dataset, each containing 1,500 tweets and with the same label distribution.

4.2.4 Classification Features

- **N-grams**

N-gram features are widely used in a variety of tasks, including tweet classification. In the study, we used unigrams (n=1) and bigrams (n=2). Special characters and emoticons were excluded from the n-gram model. We removed stop words and used the root form (lemma) of the words.

- **Text Features**

We used some structural features associated with the tweet content, such as the length of the tweets in terms of the number of words and characters, number of special characters, and Parts-Of-Speech (POS) tags. URLs were not considered in this processing. With empirical observations, we found that the tweets that contain the least number of words or more special characters were generally not informative tweets. POS features have been proven effective in tweet classification tasks.

- **Author Features**

As mentioned in the introduction, sometimes *who has said it* (the author of the tweet) can matter as much as or even more than *what has been said* (the tweet content). Twitter implements a follow-follower networking scheme in which a user can follow (subscribe to) multiple Twitter accounts as per his interests as well as have multiple followers who are interested in subscribing to his Twitter updates. The authority of the author of the tweets as well as the social networks (e.g. follower-follower relationship) of the author usually plays an important role in demonstrating the informativeness of the tweets [Yajuan et al. 2012]. We used the following features associated with the tweet authors: 1) social connectivity, i.e., number of follow-followers, 2) Twitter activity, i.e., number of tweets, and 3) authors credibility or influence, i.e., Klout score ⁶.

- **Popularity Features (social share)**

⁶<https://klout.com/corp/score>

Classifier	Precision
Naive Bayes	80.93
LibSVM	78.37
Random Forest	81.54

Table 4.2: Performance of different classifiers in the informative tweet classification task

Each tweet from the experimental dataset, contains a unique URL. One aspect to consider for tweet informativeness is the popularity of the URLs on social networks, which is measured by the level of attention they receive in the form of social shares and likes. We used the following popularity features associated with the tweets and URLs: number of retweets, Facebook shares, Facebook likes, Facebook comments, Twitter shares (tweets), and Google Plus shares.

- **PageRank**

We also used the Google PageRank of the URLs in the tweets as a feature. The PageRank algorithm has been widely used to rank web pages as well as people based on their authority and influence.

4.3 Experiments and Evaluations

We used classification precision as an evaluation metric. We performed multiple experiments with different machine learning classifiers (Naive Bayes, Random Forest, Libsvm) and different combinations of the features. Based on the experiments, we selected a Naive Bayes classifier as it was very fast (a crucial factor for classifying millions of tweets in a timely manner) and had competitive performance with respect to the other classifiers.

The following (Table 4.3) shows the summary of the experiments with Naive Bayes classifier and different combinations of the features.

As shown in the table, after using rule-based filtering, the baseline performance with tweets'

Features	Precision
Tweet	66.2
Tweet + URL Title	68.72
Tweet + URL Title + URL Content	74.67
Tweet + URL Title + URL Content + Tweet Length	74.92
Tweet + URL Title + URL Content + Tweet Length + Number of words	75.79
(Tweet + URL Title + URL Content + Tweet Length + Number of words + Special chars) =>FT1	76.83
FT1 + POS tags	77.23
FT1 + POS tags + PageRank	80.63
FT1 + POS tags + PageRank + social share	80.66
FT1 + POS tags + PageRank + social share + Author Features	80.93

Table 4.3: Classification performance with different combinations of the features

unigram and bigram was 66.20%. As we added more n-gram features like unigram and bigrams from the URL title and URL content, we achieved 74.67% precision. That is, after filtering the tweets using a rule-based approach, just using n-grams from tweet (the URL title and URL content), we can classify almost 75% of the tweets correctly. Each structural feature, like tweet length, number of words in the tweets, number of special characters in the tweet, and POS tags further improved the classification performance. URL PageRank improved the performance significantly (by 3.5%), while social share and author features only marginally improved the classification performance.

4.4 Social Health Signals System

To facilitate the browsing of the informative health information shared on Twitter, we have built a system, Social Health Signals (SHS), where a user can:

- Find reliable and popular health information from Twitter for a topic aggregated all in one place.
- Ask health related questions.
- Filter the results (tweets and URLs) by semantic health categories such as symptoms, food and diet, healthy living, and prevention.
- Visualize the tweet traffic of a topic based on location.
- Access complementary static (factual) information about the disease from Wikipedia.

The SHS system process tweets in near real-time and updates results every 6 hours. SHS is built as an extension of Twitris system. [Nagarajan et al. 2009; Jadhav et al. 2010; Sheth et al. 2014; Sheth et al. 2010; Jadhav et al. 2013; Smith et al. 2012]

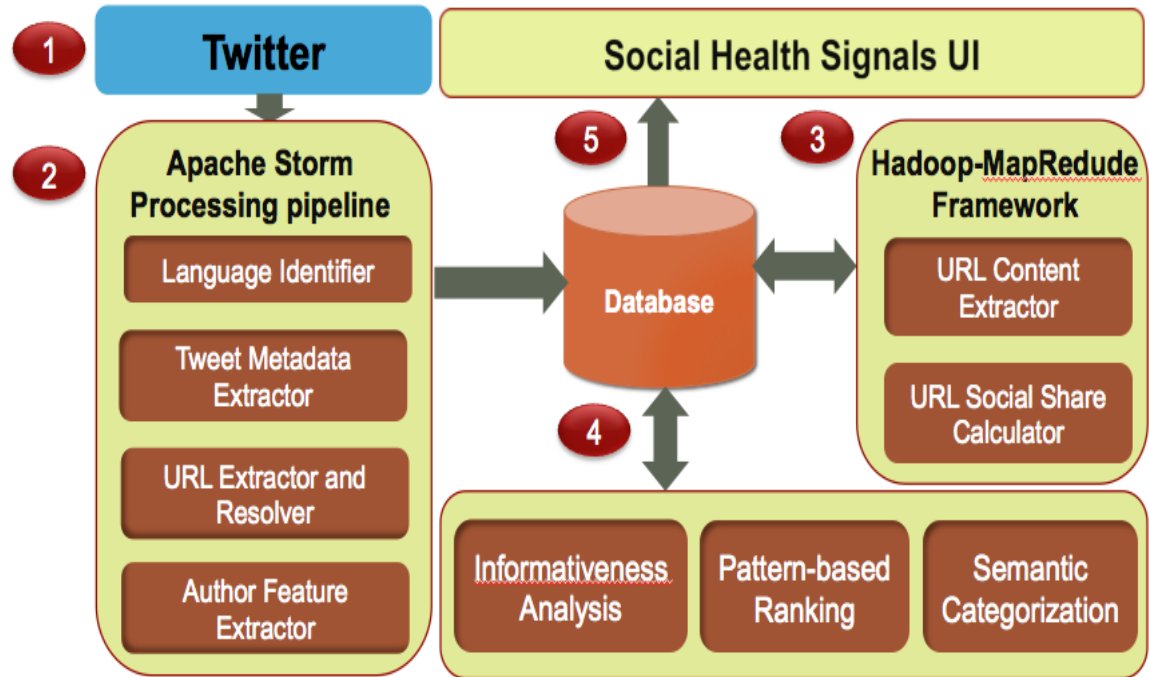


Figure 4.1: Social Health Signals Architecture

4.4.1 Data Processing Pipeline

The Social Health Signals system uses the public Twitter streaming API⁷ to collect real-time tweets related to topics (e.g. diabetes) and associated metadata. We collected keywords for the real-time crawler by studying UMLS, diabetes forums, and diabetes related tweets. We used the Apache Storm framework⁸ to perform the following analysis on the tweets:

- Language identification

To filter out non-English tweets.

- Hashtag retrieval

To retrieve hashtags from the tweets.

⁷<https://dev.twitter.com/streaming/>

⁸<http://storm.apache.org/>

- URL extractor

To extract URLs from tweets.

- URL resolver

To expand shortened URLs to their original form.

- Location retrieval

To retrieve the geo-coordinates of the tweets.

- Author feature extractor

To extract metadata about the author of the tweet, such as his follow-follower count and number of tweets.

- PageRank calculator

To calculate the PageRank of the extracted, resolved, and working URLs.

- Tweet content feature extractor

To extract the number of words, special characters, spelling mistakes, and POS tags.

- Semantic annotation

To annotate tweet content and URL title with UMLS concepts and Semantic Types.

SHS also uses a Hadoop-MapReduce framework to extract URL content and a URL's social share (on Twitter, Facebook) count. Finally, SHS classifies the informative tweets using the classification algorithm described in the previous section.

4.4.2 Question and Answering on Twitter data

One of the features of SHS system is to let users to ask health-related questions on Twitter data. SHS ranks the informative tweets and URLs using Annotated Query Language (AQL)-based patterns for question-answering [Soni 2015]. AQL is a query language that extracts structured information from

unstructured or semi-structured text. We have used triple-based (subject, predicate, and object) pattern mining technique to extract triple patterns from tweets and users' questions. For user query expansion, we have incorporated the domain knowledge using the UMLS Metathesaurus and WordNet. Once the relevant results for a user query are retrieved, SHS uses a Random Forest classifier to rank the results based on the social share and relevancy (with the user query) features of the results.

4.4.3 Semantic Categorization

In the data processing, all the informative tweets and URLs' titles are annotated with UMLS concepts and Semantic Types. To enable efficient browsing of the health information, SHS uses a search intent mining algorithm (Chapter 3) which classifies informative tweets and URLs into consumer-oriented health categories like Symptoms, Living with, Food and diet, Prevention and Treatments. Such categorization enables users to further filter the informative tweets by health categories of their interest. For example, if a user is interested in the prevention-related information, then once the user selects prevention in the SHS "Top Health News" interface, only prevention-related tweets and news articles will be shown.

4.4.4 Social Health Signals User Interface

Figure 4.2, shows the user interface of the Social Health Signals system. Following are the major components of the SHS UI.

- Search and Explore

To ask questions or perform search on informative tweets and URLs.

- Top Health News

List of informative URLs based on URLs extracted from informative tweets.

Can be filtered based on health categories.

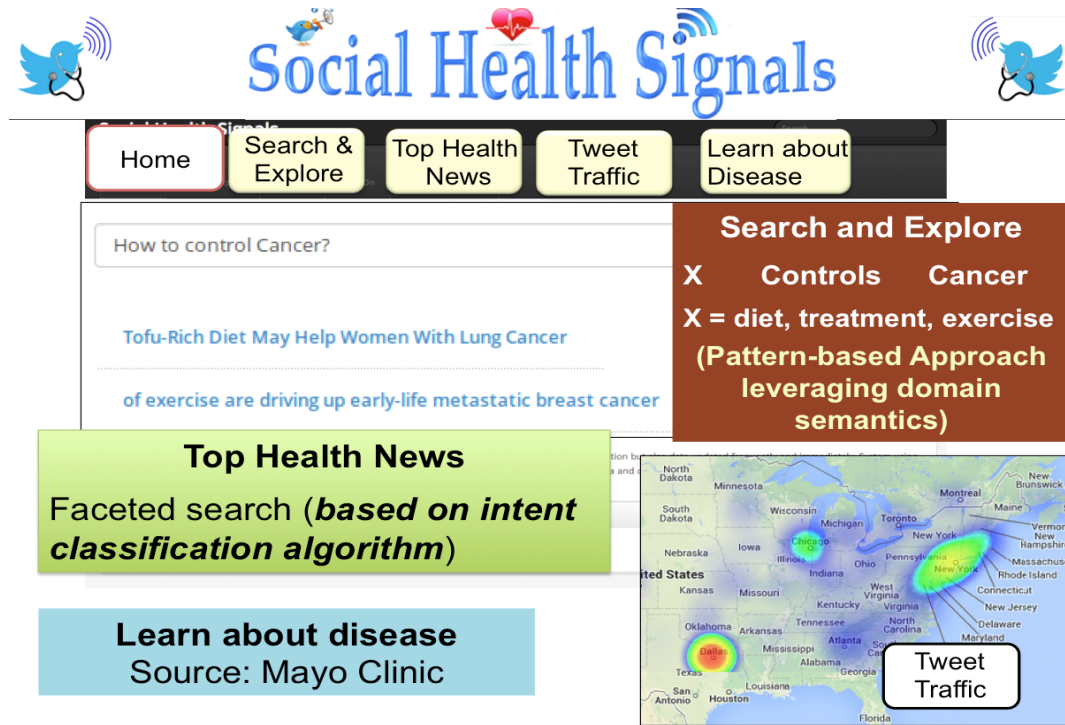


Figure 4.2: Social Health Signals User Interface

- Tweet traffic
 - Heatmap of location-based distribution of tweets related to a topic.
- Learn about Disease
 - Static information about diseases from Wikipedia and Mayo Clinic.

5

Evaluating the Process of Online Health Information Searching: A Qualitative Approach to Exploring Consumer Perspectives

In the Chapter 3, we briefly described the focus group study that we conducted to understand consumers' perspective on online health information seeking. In this chapter, we will present this study in detail.

5.1 Introduction

In recent years, the quantity and quality of health information available on the Internet has improved substantially. As access to reliable, affordable, high-speed Internet access increases, the percentage of people using the Internet to search and subsequently learn from health-related information continues

to grow rapidly as well. In the current climate of rising costs of health care in the United States, the role of freely available health care information is becoming more central to patients, their families and friends, and even health care providers. In order to improve the delivery of content, researchers and scientists must first develop a thorough understanding of the searching-related needs and experiences of users.

Recent studies have shed light on why and how consumers search for health information on the Internet [Fox 2014], [Gutierrez et al. 2014] [CHAUDHURI et al. 2013]. In a recent 2013 survey conducted by the Pew Internet Project, 72% of respondents reported using the Internet to look for health information within the past year, with the most commonly researched topics being focused on specific diseases or conditions, treatments or procedures, and searching for doctors or other health professionals [Fox 2014]. Although many people (35% of those surveyed by Pew) use the Internet to learn more about a specific symptom or medical condition they or someone else might have, clinicians and/or family and friends remain a central resource when help is needed regarding a serious health issue [Gutierrez et al. 2014], [Cutilli 2010]. The elderly in particular are more likely to trust “living sources” of information, rather than the Internet [CHAUDHURI et al. 2013]. Even among Internet users, health information is often understood in a social context. For example, 26% of Internet users reported watching or reading content related to someone else's personal experience with a medical or health-related issue within the last 12 months [Fox 2014].

Health information seeking behavior depends on a variety of factors including subjective factors (e.g., intent for the search, experience in using and searching the Internet, and information preferences [Higgins et al. 2011], [Lorence et al. 2006]) and socioeconomic factors (eg, age group, income level, education level, etc [Cutilli 2010], [Drentea et al. 2008], [Flynn et al. 2006]). Research shows that women are more likely than men to search for health information [Atkinson et al. 2009] and online health consumers tend to be more educated, earn more, and have high-speed Internet access at home and at work [Higgins et al. 2011], [Wangberg et al. 2008], [Kummervold et al. 2008]. Although low-income individuals do use the Internet, some may have difficulty distinguishing be-

tween low and high quality information [Knapp et al. 2011]. Additionally, low-income disabled and homebound adults show lower rates of Internet use overall [Choi and DiNitto 2013]. Further, our preliminary results from another study indicate that online health information seeking behavior differs significantly compared to general information searching. In particular, our data suggests that health-related queries are typically longer (ie, more words) and contextual in nature compared to general queries [Jadhav et al. 2014]. Also, health-related queries have higher rates of misspelled words that are typically corrected by “auto-completion” features available universally in all Web search engines such as Google and Bing [Jadhav et al. 2014].

There are various motivating factors for health information searching on the Internet. Aside from trying to learn more about a symptom or disorder specifically relevant to the person searching, half of online health information research is on behalf of a friend or relative [Sadasivam et al. 2013]. Additionally, searching is often used to track specific health-related factors. For example, 60% of adults reported tracking their weight diet or exercise routine online, and 33% reported tracking specific health indicators or symptoms such as blood pressure, blood sugar, headaches, or sleep patterns [Fox 2014].

A large proportion of the population uses the Internet to search for health information, and their motivations for doing so are varied, [Gutierrez et al. 2014] [CHAUDHURI et al. 2013]. This complex situation, along with an educationally and culturally heterogeneous population, has resulted in a barrier in the process of gathering and interpreting health information. In this context, the preferred vocabulary within and between different groups of people can differ significantly, often resulting in a variety of words being used to describe the same concept or medical condition [Smith 2007], [Keselman et al. 2008], [Zielstorff 2003]. Knowledge gaps can then emerge between patients and providers. One possible strategy for addressing such gaps involves developing consumer-focused vocabularies and associated infrastructure for health information retrieval that can act as an interface between parties [Seedor et al. 2013]. Before such vocabularies and technologies can be developed, researchers and scientists must have a thorough understanding of the current state of online health

information searching. While a large body of survey-based research has been conducted regarding this subject [Fox 2014], [Gutierrez et al. 2014], [Koch-Weser et al. 2010], qualitative research provides a unique perspective that can play a valuable role in informing future research and technological developments. In this study our objective was to engage in in-depth discussions with community members about their health-related searching activities. All the study participants are residents of Olmsted County, Minnesota (MN), and are either Mayo Clinic patients, employees, or at least have one family member at home who is a patient or employee.

5.2 Methods

5.2.1 Study Participants and Recruitment

To better understand health information searching behavior and its implications for health and well-being of community members, we conducted three 90-minute focus groups of 5 to 6 individuals over the course of a 2-month span. We targeted adult, English-speaking members of the Olmsted County, MN community (where Mayo Clinic is located) and Mayo Clinic patient, employees, and family visitors. We recruited participants using flyers and online classifieds ads distributed throughout the Rochester, MN community and within Mayo Clinic. Table 1 summarizes basic characteristics of participants. Participants were provided a modest financial remuneration for participating in the study.

Moderators (JM and AK) trained in qualitative methodology facilitated discussions about the attitudes and experiences of participants related to searching for health information on the Internet. Moderators used a semi-structured moderator guide to facilitate discussion and the guide covered four major aspects: (1) participants perception and understanding of health care information, (2) the process of information collection on the Internet, (3) understanding and usage of information, and (4) implications of health care information for their health and well-being. Participants were asked about their thoughts and the connotations surrounding each of these themes. Oral consent

was obtained from all participants. This study was approved by the Institutional Review Board at Mayo Clinic (IRB #12-005476).

Prior to participating in the focus groups, participants completed an anonymous questionnaire that included questions assessing basic demographic information and previously used sources of health information. All focus groups were audio-recorded, transcribed, and de-identified.

5.2.2 Data Collection and Analysis

All team members read de-identified transcripts and developed a codebook through an iterative process [Corbin and Strauss 2014]. Using the codebook, two members of the team independently coded the transcripts in NVivo, a qualitative software application. The data were then analyzed using a grounded-theory approach (NVivo qualitative data analysis software; QSR International Pty Ltd. Version 10, 2012). Coding inconsistencies were discussed and resolved through consensus, with the input of a third team member when necessary.

5.3 Results

5.3.1 Overview

Participants candidly discussed how they used the Internet to search for health information. Through these discussions, several themes related to health motivations, content preferences, and practical applications of searching emerged. Below we summarize this data in the context of three major themes: motivations for searching, searching strategies and techniques, and information content preferences.

5.3.2 Motivations for Online Health Searching

A variety of factors play a role in initiating online searches for health information. The motivations that our participants described generally fell into three main areas: (1) symptom troubleshooting,

Characteristic		n (%)
Age		43.26 (17.0; 22-73)
Sex	Male	5 (26%)
	Female	14 (74%)
Race	White	15 (79%)
	Black or African American	0 (0%)
	Asian	4 (21%)
Highest level of education	High school or GED	0 (0%)
	Community or Jr. College	3 (16%)
	Four-year college	3 (16%)
	Graduate school	13 (68%)
Yearly household income (US\$)	Less than \$15,000	0 (0%)
	15,000–35,000	2 (11%)
	35,001–55,000	9 (47%)
	55,001–75,000	4 (21%)
	75,001–100,000	0 (0%)
	Over \$100,000	1 (5%)
	Prefer not to answer	3 (16%)
Prior sources used to get health information	Health care providers	19 (100%)
	Family/friends	15 (79%)
	Organizations/support groups	6 (32%)
	Internet	18 (95%)
	Books/pamphlets	15 (79%)
	Other	1 (5%)
Prior participation in research	Yes	4 (21%)
	No	15 (79%)

Table 5.1: Characteristics of patients (n=19).

(2) searching to enhance a clinic visit, and (3) proxy searching.

Perhaps the most common motivation for everyday searching is a phenomenon that could be called “symptom troubleshooting”. With commercial online resources and other government or hospital/university-based sites that provide free, anonymous, and immediate information, many individuals' first stop to learn more about a specific symptom is the Internet. A participant from Focus Group (FG) #3 mentioned: “For me, it was very important when I think I have a symptom, the first place I look is the Internet, especially to search for the symptoms”.

Once a particular symptom or disorder of interest is identified, participants reported that the Internet made it very easy to get more detailed information to help identify underlying causes. As a participant of FG #3 explained: “For instance if I have a pain in my foot, I am going to start looking for... information that might specify if it's in the heel or in the toe... then I search [for] why [I have] the symptom or, if I know what I have, then I might search... to see if I can match the symptoms to that”.

Using the Internet provided a quick and easy way to troubleshoot symptoms; however, there are certain situations where using the Internet is more likely. One participant explained that the Internet is especially more convenient for superficial symptoms: “You can't just go find a doctor somewhere and be ‘hey, can you look at this rash on my leg’ because I hear doctors hate that” [FG #1]. The Internet provides a level of anonymity that may be helpful in situations where individuals perceive their problems to be bothersome or nuisances to doctors. Participants often cited practical reasons related to time and money when describing their motivations for turning to the Internet for medical information or advice. One participant explained that although consulting a professional in person can be preferable, “especially when you are very concerned about your symptoms”, in other cases, as he stated, “at 9:00 at night you are not going to be able to call the doctor” [FG #3]. Another participant in FG #1 also echoed a similar sentiment: “It can't be readily available, you may have to make a doctor's appointment and that could take a while... and cost money and financially that might hold you back too; something that a fast care isn't going to be able to fix”. For non-serious

medical issues, participants were generally comfortable using the Internet as a troubleshooting tool. Once a health care provider is involved, however, searching assumes a different role. In this context, participants reported using Internet searching as a means to enhance a clinic visit and be more well-prepared and well-informed during the entire health care experience with their providers. In these situations, Internet searching proved to be a valuable tool in preparing for the clinic visit. As one participant in FG #1 explained, Internet searching allowed her to walk into a surgery consultation armed with a prior understanding of possible procedures: “I specifically knew all the three main surgeries; I knew what I liked from them, what I didn't like of them”.

This online preparation gave her the information and ability to “say what about this, what about that, why are we doing this, why are we doing that?” [FG #1]. Participants agreed that such preparation facilitates “a more enriched experience” [FG #1] and allows patients to “become more knowledgeable” and “ask better questions” to providers [FG #2]. This participant goes on to explain how such a dynamic increases communication and education and “builds the patient/provider relationship”; “If you are taking an interest in what it is you have and asking the kind of questions that allow them to further educate you, I think that shows a real interest” [FG #2].

Another participant expanded on this idea and explained how an enriched patient/provider relationship involves more than developing a healthy rapport and can actually improve health outcomes in certain situations: “I mean my mom had a weird thyroid thing and she was all over the Internet, and still is, but she would bring stuff to her doctor and she actually like did solve some mysterious things and she gave stuff to her doctor and her doctor I think is a great doctor but there is so much information and the doctors don't get it all” [FG #2].

In the previous example, the participant's mother used the Internet for two of the main motivations that emerged from our focus groups: to troubleshoot a thyroid condition and to enhance her visits with her doctors. Although this participant's mother was able to do the searching and advocating on her own, many participants had parents, grandparents, or other family members who were not as comfortable or capable. These situations highlight the third main motivation for

searching that our participants discussed: searching for someone else, or proxy searching. All of the focus groups had participants who reported searching on behalf of someone else. For many, it was a frequent occurrence.

Computer literacy was often cited as a main reason for proxy searching, as many participants had relatives who were “afraid of using it [computers and the Internet]” [FG #1]. However, proxy searching was also a useful tactic when the individual searching sought to protect their relative from additional emotional burdens, even when the relative was computer literate. One focus group participant explained: “Well, I have done searches for my parents before. . . When I looked up stuff [about] breast cancer on the Internet, [I told them] do not look it up because you're going to be scared. As a third person, even though she is my mom, I know how to decide and to remove myself from the situation, but she is not going to be able to do that” [FG #3].

5.3.3 Searching Strategies and Techniques

In terms of the actual mechanics of searching, participants described using a common set of steps and procedures that began with commonly used search engines, continued to shop around for information from various sources, and ended with information saturation and exhaustion.

Regardless of the underlying motivations for searching, almost all searches shared a common starting point from an online Web search engine: Google. Ease of use “you can ask the most stupidest questions and have a pretty good shot of getting an answer” [FG #1] and quality of results “[Google] brings up the most variety of answers” [FG #1] were the primary reasons for choosing Google cited by our participants.

Although Google is by far the most common first step to searching, its main use is simply as a tool to reach other sites. One participant mentioned: “Google's just a way to get there” [FG #1]. Another participant expanded on this view, adding “I agree. I am not putting my trust in Google; I am only putting my trust that Google is going to give me a variety. My trust is actually embedded only in the searches I click, it is just the outlet to get me there, it is just the bridg” [FG #1].

Once Google supplied a list of relevant sites to visit, most participants reported visiting many sites in order to satisfy their searching demands. This technique allows participants to “shop around and have multiple sources” without having to use exact phrasing [FG #1]. The information shopping process described by participants often included multiple side-by-side comparisons. One participant mentioned: “Because you can multiple open window task bars and tabs on the Web browser, I open every single one on the first page in each of the task bars and compare all of them” [FG #2].

This technique facilitated the information shopping experience and gives greater confidence in results because “you get as much information as you can if [all the websites] have the same information” [FG #2]. Many participants used the tabs function of Web browsers to compare multiple websites at once.

Participants described a common sequence of events that led to the termination of the search process. As the comparing and filtering process of multiple websites progresses, participants reported that eventually “all the information is basically the same” [FG #1]. Although another participant acknowledged that “there are always additional links to go to” [FG #1], other participants explained that once results became irrelevant to their original search query it was time to stop the search process. One participant explained: “If you go down to the 17th, 20th, 30th option under Google, you find that what you are looking for is the 30th degree of separation. It is just not as relevant to what it is you are trying to research anymore” [FG #1].

Some participants also reported a sense of being “lost” or “completely forgetting where you started”, especially in cases of performing broad searches. The resulting confusion can lead to becoming “unmotivated” to continue searching, even if the original query has not been resolved [FG #1].

In addition to information saturation, subjective fatigue was an indicator participants described as a reason for ending the search process. After a long, drawn-out search process, participants reported getting “tired with the screens” and feeling “exhausted” [FG #1]. Another participant compared the process to shopping: “If you know what you want, you can go to ten different places

to try to find that one thing, but after a while...you are going to be hitting your head against the wall...it gets exhausting" [FG #1].

Ultimately, the participants described searching for health-related information as a rigorous process of comparing and contrasting various sources against personalized criteria based on need and individual appraisal of reputation. This filtering process generally continues until the results become repetitive and/or the searcher becomes fatigued.

5.3.4 Content Preferences

Major search engines can easily produce thousands of results for any given query. How then do patients and consumers select which websites to gather health-related information? Although every search is unique, participants overwhelmingly preferred sites based on two main factors: reputation and advertising (or lack thereof).

Participants often commented that they "tend to go for the sites that are most reputable" [FG #1]. While the importance of reputation applied to all websites, regardless of if they were related to health, participants also reported placing a higher standard of quality on health-related information. As one participant explained, "Health is unlike any other consumer type of website...I take it to a totally different level. I want to have the best, you only have one body" [FG #1]. Making sure they had "the best" gave participants comfort in knowing they were receiving accurate information. Often the best is synonymous with dealing with a "reputable institution", which is in turn largely influenced by branding. One participant explained: "When you are dealing with a company, an organization that has a good reputation, then you feel more confident that you are getting the right information" [FG #3].

In addition to pure name recognition, participants reported that institutions "earn trust... through publications, research, and education" [FG #2]. Additionally, "how [websites or institutions] are ranked" or if they are "well known" contributed to participants' conception of reputation [FG #2]. Finally, participants were more likely to view sources of health information as reputable if they were

domestic. As one participant explained, “I would rely more heavily on those [domestic] institutions than a foreign hospital that may be quite good but is somewhere outside of the United States” [FG #2].

While reputation played a major role in determining which websites to trust for our participants, advertising and commercial interests often dissuaded them. Almost all of our participants reported avoiding websites that had visible advertising or were obviously profit-oriented. As one participant explained, “If I see ads, I question the motivation for providing information that they have” [FG #1]. Another participant explained the aversion in the context of a wider trend of commercialization of medicine: “I think for me it scares me how, and I suppose this could go onto a variety of different things, but it scares me how medicine has transformed into such a consumer-driven place” [FG #1].

Most of our participants shared distaste for commercial interests in their searching behavior; however, in some cases it had more to do with the perception of profit-driven motivations rather than the true nature of the business or organization. In response to a question regarding whether or not participants thought that MayoClinic.com, the commercial consumer health information portal owned and maintained by Mayo Clinic, was a “commercial” website, one participant responded, “Well, you don't see a lot of advertising on the Mayo site . . . I don't see a lot going on the sides all the way down the page flashing at me, I don't have a lot of popups that come at me” [FG #1].

Although Mayo Clinic does indeed utilize advertising on the website, the combined name recognition, familiarity, and subtle nature of advertisements was enough to retain credibility for many of our participants. We acknowledge that there might be an inherent bias in this finding since the study participants were either Mayo Clinic patients, employees, or at least have one family member at home who is a patient or employee.

5.4 Discussion

5.4.1 Principal Findings

Our goal in collecting these qualitative data was to better understand how consumers use and search for health information on the Internet to inform the development of more personalized health information searching and delivery applications. The participants in this study described a common experience of searching for health information that largely mirrors recent large-scale survey data. Most of our participants see the Internet as a potentially valuable tool to find information about health and medical conditions; yet, they did point to the challenge of efficiently addressing their particular needs given the vast amounts of information. This reflects the challenge of streamlining and personalizing information for a user base that is diverse both in terms of individual background and need. The data presented here, particularly in the context of content preferences and searching techniques, may be beneficial to researchers and content providers as they develop new strategies for delivering health information.

Many participants shared examples of how they use information they found through Internet searches in their efforts to enhance their interactions with their health care providers. Examining these data in the context of increasing health costs and physician time constraints provides valuable insight into the challenges and opportunities consumers and physicians will encounter in years ahead. Many of our participants reported using Internet health searching as a means of enhancing clinic visits, either through preparation or post-appointment follow-up. Some concerns exist regarding how doctors may react to patients introducing health information gathered from the Internet into the exam room, and indeed previous research has indicated that some physicians view such occurrences negatively [Hamann et al. 2012] [Ahmad et al. 2006]. Patients, on the other hand, tend to view Internet health searching as an additional resource to complement the still highly valued patient/physician relationship [Kivits 2006], [Stevenson et al. 2007]. Our data also support this

view of the patient perspective, as our participants viewed online health searching as a means to “build the doctor-patient relationship” [FG #2]. How physicians respond likely depends on physician communication skills and whether or not the physician feels challenged [Murray et al. 2003]. The participant experiences and opinions described here are largely from a patient perspective and are largely positive in the context of using health information from the Internet to enhance visits. These perspectives may be useful in framing future research focused on physician perspectives on using such information in office visits.

Recently, the amount of time doctors spend in front of patients has received attention in the media [Chen 2013], [Block et al. 2013]. Having patients armed with information and questions prior to office visits may help improve care in the current realities of decreased face time with doctors, which today can be as low as 8 minutes on average [Block et al. 2013]. This of course necessitates that the information patients gather be of high quality. Indeed, research suggests the quality of information that patients present ultimately determines its effect on the patient/physician relationship; while accurate information can be helpful, inaccurate information may be harmful [Murray et al. 2003]. Our future work will therefore focus on ways to develop consumer health information technology solutions to facilitate the transmission of accurate, trustworthy, validated information to consumers to ensure that online health information searching enhances, rather than hinders, care.

5.4.2 Limitations

This study contained a few important limitations. Due to recruitment constraints, our study population was limited to adults within Olmsted County, MN. All participants were either employees or were family members of employees and patients at Mayo Clinic, where the study took place. Additionally, our sample was highly educated, with all participants having attained at least a community college degree, and 68% having completed graduate school. We were therefore unable to explore the perspectives of a more diverse population. It is also important to consider our choice of study design when interpreting the data we presented. In this study, we used qualitative approaches such as

grounded theory and focus groups method for data collection and analysis. These qualitative methods allow us to contextualize participants understandings and experiences, to track variations in how concepts are understood, and to uncover novel findings that may warrant further investigation [Sulmasy and Sugarman 2001]. In this way, we are able to make, as Giacomini and Cook describe, an “empirically-based contribution to ongoing dialogue” [Giacomini et al. 2000]. The overarching goal of qualitative research is to explore and describe particularities of a social phenomenon rather than producing generalizable results. But, findings from a small sample size in a qualitative research can help developing hypothesis for a quantitative study to produce generalizable findings from a larger sample size. Our study participants were recruited from a limited subset of individuals that was readily accessible in a community dominated by the health care industry. In doing so, our goal is not to present data that can or should be generalized to a wider population, but rather to explore pertinent issues with a level of depth that is not possible with standard quantitative (and generalizable) methodologies. Indeed, we cannot claim that the experiences described here are representative of all Internet users; however, they can inform the development of future work and research in areas of streamlining content delivery and patient/physician interaction.

5.5 Conclusion

We conducted this qualitative study to gain a deeper understanding of search behavior in order to inform future technological developments in personalizing online information searching and content delivery. This study provided important insights and helped us to understand:

- Consumers' perspective (e.g. their experiences, challenges) about online health information seeking.
- Why (motivations) and how (search strategies) participants use the Internet to seek for health information.
- What health information (intent classes) do they search using the Internet.

Although the Internet was a preferred source of health information for almost all of our participants, from a consumer and patient perspective challenges persist in streamlining the process of identifying reliable and high quality content that also matches the intended search target of the user. Our participants described a current search paradigm consisting of drawn-out user-driven comparisons of content obtained from multiple sources of varying quality and unverified validity. As consumers continue to use information gathered from the Internet to enhance their interactions with health care providers, new strategies for delivering health information on the Internet must be developed that accommodate diverse backgrounds and clinical needs.

6

Comparative Analysis of Expressions of Search Intent From Personal Computers and Smart Devices

In the previous chapters, we covered topics related to the identification of search intents from the Web search queries. In this chapter, we will compare expression of health search intents and associated features by analyzing large-scale health related search queries generated from desktop devices and smart devices.

6.1 Introduction

With the recent exponential growth in usage of smart devices (SDs) like smartphones and tablets, the percentage of Online Health Information Seekers (OHISs) using smart devices to search for health

information has grown rapidly [Duggan and Smith 2013], [Fox and Duggan 2012]. In 2015, Google revealed¹ that more Google searches take place on smart devices than on personal computers in 10 countries including the US and Japan. While there is some evidence [Roto 2006] that the experience of online information searching varies depending on the device used (eg, smart devices vs personal computers or laptops [PCs]), little is known about how device choice impacts the structure of search queries generated by users. Understanding the effects of the device used (SDs vs PCs) for health information search would help us to acquire more insights into online health information seeking behavior (OHISB). Such knowledge can be applied to improve the search experience and to develop more advanced next-generation knowledge and content delivery systems. In this study [Jadhav et al. 2014; Jadhav and Pathak 2014], we compare health search intents and features that contribute to the expression of search intent by analyzing large-scale health related search queries generated from PCs and SDs.

Using the Mayo Clinic website's Web analytics tool (IBM NetInsight OnDemand²) and based on the type of devices used (PCs or SDs), we obtained the most frequent health search queries submitted from Web search engines that direct traffic to the Mayo Clinic³ webpages. We selected search queries that are in the English language and collected between June 2011 and May 2013. We analyzed structural properties, types (keywords, wh-question, yes/no-questions), misspellings, and the linguistic structure of the health queries. We further categorized them based on health categories and demographic information mentioned (gender, age group, etc) in the queries. Our analysis suggests that the device used for online health information searching plays a significant role, altering the OHISB.

¹<http://adwords.blogspot.com/2015/05/building-for-next-moment.html>

²<http://www-03.ibm.com/software/products/en/on-premise-web-analytics>

³<http://www.mayoclinic.org/>

6.1.1 Significance of Current Study

Many previous studies have investigated OHISB. Researchers have used several approaches to understand OHISB including (1) focus groups and user surveys [Lorence et al. 2006; Drentea et al. 2008; Kummervold et al. 2008; Weaver III et al. 2010; Eysenbach and Köhler 2002; Wangberg et al. 2008; Atkinson et al. 2009] and (2) analyzing health-related Web search query logs [White and Horvitz 2014; 2009a; 2009b]. In the studies that involved focus groups and user surveys, researchers have analyzed characteristics associated with OHISB such as how people use the Internet for health information searching, their demographic information (age, gender, education level, etc), devices/Web search engines used for searching, OHISB in specific health conditions, and age groups [Wangberg et al. 2008; Atkinson et al. 2009; Flynn et al. 2006]. Although these studies provide important insights into OHISB, their main limitation is the inclusion of a small number of participants (ranging from 100-2000 people). A second approach to studying OHISB is analyzing Web search logs from the health domain. Several previous studies have analyzed health search query logs with diverse objectives, such as health/epidemic surveillance [Ginsberg et al. 2009; Ocampo et al. 2013; Brownstein et al. 2009; Carneiro and Mylonakis 2009], PubMed usage [Herskovic et al. 2007; Dogan et al. 2009], and OHISB [White and Horvitz 2014; 2009a; 2009b; Zhang et al. 2012; Eysenbach and Köhler 2004]. The studies focusing on OHISB [White and Horvitz 2014; 2009a; 2009b; Zhang et al. 2012; Eysenbach and Köhler 2004] have considered a variety of aspects of health query logs, such as query length, health categories, relationship between OHISB and health care utilization [White and Horvitz 2013], changes in health behavior with type of disease [White and Horvitz 2014], and changes in OHISB with disease escalation from symptoms to serious illness [White and Horvitz 2009a; 2009b].

Although the user experience for online information searching varies with the device used (PCs/SDs) [Roto 2006], there is a dearth of work relating OHISB with the device used for searching. In this study, we address this problem by analyzing large-scale health queries for both PCs and SDs to

understand the effects of device type (PCs vs SDs) used for online health information seeking. Previous studies in generic search query log analysis have determined the importance of understanding linguistic structure of search queries as it has implications on information retrieval using Web search engines [Barr et al. 2008; Croft et al. 2010]. One of the contributions of this study is a comparative analysis of linguistic structure of health search queries from PCs and SDs. This study provides useful and interesting findings that can be leveraged in multiple ways. Some of the potential beneficiaries are (1) **Web search engines**: to understand health search query structure and complexity, and the occurrence of popular health categories for PCs and SDs to improve query performance and accuracy for health information retrieval systems, (2) **Websites that provide health information**: to better understand online health information seekers' health information need, and better organize health information content for PCs and SDs users, (3) **Healthcare providers**: to better understand their patients and their health information interests, (4) **Health care-centric application developers**: to better understand OHISB for PCs and SDs and build applications around consumer's health information needs and priorities, and (5) **Online health information seekers**: to empower online health information seekers in their quest for health information and facilitate their health information search efforts by enabling the development of smarter and more sophisticated consumer health information delivery mechanisms.

6.2 Methods

6.2.1 Data Source

In this study, we collected health search queries originating from Web search engines (such as Google and Bing) that direct OHISs to the Mayo Clinic's consumer health information website⁴, which is one of the top online health information website within the United States. The Mayo Clinic website is identical in terms of appearance and functionality for both PCs and SDs using standard Web search

⁴<http://www.mayoclinic.org/>

engines and Web browsers. This consistency as well as significant traffic to the website provide us with an excellent platform to conduct our study.

6.2.2 Dataset Creation

The Mayo Clinic website's Web analytics tool, IBM NetInsight OnDemand⁵, keeps detailed information about incoming Web traffic from Web search engines to the Mayo Clinic website. The tool maintains information such as input search query (the original query from a Web search engine that brings an OHIS to the Mayo Clinic website), number of query repetitions (how many times the query has been searched within specified time period), and the visitor's Operating System (OS). PCs generally use Windows (98, 2000, Xp, Vista, 7, 8), Mac OS X, or Linux (such as Ubuntu and Redhat) operating systems while SDs use iOS (iPhone's OS), Android, Windows Mobile, and RIM BlackBerry operating systems. Since the Web analytics tool tracks information related to each user's OS type and individual searches, we are able to differentiate search queries by device type (PCs/SDs).

Using the Web analytics tool, we obtained one data report for each of the most frequent one million (based on the number of query repetitions) anonymized distinct queries in the English language launched from PCs and SDs for each month between June 2011 and May 2013 (24 months), totaling 48 data reports. Each search query appears uniquely in each data report and has an associated number of query repetitions. For each device type (PCs and SDs), we aggregated 24 reports to create a single report with distinct queries. The dataset for PCs has 2.74 million queries, and the dataset for SDs has 3.94 million queries. While aggregating the search queries for PCs and SDs, we combined the repetition counts for each repeated query; for example, if a “diabetes” query has 5 repetitions in 1 month and 10 repetitions in another month, then the total number of repetition for the “diabetes” query is 15. Note that selecting the top queries for 2 years would be an easier approach for dataset creation, but in our case the data reports were available by month, thus we

⁵<http://www-03.ibm.com/software/products/en/on-premise-web-analytics/>

have to aggregate the data for each month to create the final analysis dataset.

6.2.3 Data Analysis

In this study, we performed analyses on “queries with considering repetition counts (QwR)” and “queries without considering repetition counts (QwoR)”. Because the analysis performed with only QwR may overrepresent certain queries due to their large number of repetitions, we performed the analysis for both QwoR and QwR. The QwoR count is the same as the number of queries in the dataset. Hence for PCs, we have 2.74 million QwoR, and for SDs we have 3.94 million QwoR. We obtained the QwR count by aggregating number of repetitions for all the queries in the dataset. For both PCs and SDs, we got more than 100 million QwR. Due to Mayo Clinic's confidentiality policy, we are not able to disclose the exact number of QwR. We are reporting percentages of PC and SD queries.

Top Health Queries

The top search queries are the most commonly searched queries. To analyze the top health queries launched from PCs and SDs, we selected the top 100 search queries, from PCs and SDs, based on the descending order of number of query repetitions in the analysis dataset.

Health Categories

To analyze popular health categories that OHISs search for from PCs and SDs, we selected the following 8 health categories corresponding to the organization of health topics on popular health websites (Mayo Clinic, MedlinePlus⁶, WebMD⁷): Symptoms, Causes, Complications, Tests and Diagnosis, Treatments and Drugs, Risk Factors, Prevention, Coping and Support. For example, Figure 6.1 shows different health categories for diabetes on the Mayo Clinic website, where each health category has a separate webpage with detailed information (browsable via navigating the left panel).

⁶<http://www.nlm.nih.gov/medlineplus/>

⁷<http://www.webmd.com/>

The screenshot shows the Mayo Clinic website for Diabetes. The navigation bar includes 'Patient Care', 'Health Information', 'MAYO CLINIC', 'For Medical Professionals', 'Research', and 'Education'. Below the navigation bar is a search bar and a menu with categories like 'Diseases and Conditions', 'Symptoms', 'Drugs and Supplements', 'Tests and Procedures', 'Healthy Lifestyle', and 'First Aid'. The main content area is titled 'Diabetes' and includes a sidebar with a 'Definition' section highlighted by a red box. A red arrow points from the text 'Health Categories' to the 'Definition' section. The main text describes diabetes mellitus and its symptoms. There is also a 'Mayo Clinic Store' advertisement on the right.

Figure 6.1: Screenshot of Mayo Clinic website for Diabetes (left-side box highlights organization of health information based on health categories)

Based on the semantics of an OHIS's input search query and a Web search engine's recommendations, users may land on one of the health category pages on the Mayo Clinic website. For this study, we aggregated all the incoming health search queries between June 2011 and May 2013 that land on a particular health category webpages. For example, we aggregated all the search queries that land on the "Symptoms" webpage for all the diseases and health conditions on the Mayo Clinic website. We analyzed the type of device (PC or SD) used for searches and the number of search queries to each health category.

Categorization Based on the Information Mentioned in the Health Queries

In order to understand how often an OHIS mentions gender, age groups, and temporal references in the search queries, we categorized health queries using a dictionary-based approach. For each

group, we created a lexicon by going through online English dictionaries⁸ and a manual evaluation of words. For example, in the “Gender” group we considered Men (Man, men, male, boy, gent, gentleman, gentlemen) and Women (Woman, women, female, girl, ladies, lady). We also considered keywords’ lexical variants; for example, boy, boys, etc. We categorized search queries from PCs and SDs by utilizing the lexicon for each category.

Health Query Length

To study the difference in health search query length for queries from PCs and SDs, we calculated search query length by computing the number of words (separated by white space) and the number of characters (excluding white space) in the health queries.

Usage of Query Operators and Special Characters

In search queries, query operators (“and”, “or”, “not”, etc) are used to formulate complex queries. In this study, we considered the following operators: AND, OR, +, &, other (NOT, AND NOT, OR NOT, & NOT). Special characters are characters apart from letters (a-z) and digits (0-9). The significance of special characters in a health search query depends on the usage of special characters in the medical domain. For example, OHIS may mention values in different formats, eg, 2.3 ml, 40%, 17-19, or \$200 (for the cost of a drug or procedure). We analyzed the usage of search query operators and special characters in health queries based on their usage frequency in the PCs and SDs search queries.

Misspellings in Health Queries

OHISs occasionally make spelling mistakes while searching for health information. To analyze the frequency of such errors, we used a dictionary-based approach. We first generated a dictionary of

⁸<http://dictionary.reference.com/>, <http://www.merriam-webster.com/>, <http://www.oxforddictionaries.com/us>

words using the Zyzzyva wordlist⁹, the Hunspell dictionary¹⁰, and its medical version (OpenMed-Spell), comprising a total of 275,270 unique words. We used this dictionary to check misspellings in health search queries from PCs and SDs.

Type of Search Queries

OHISs express their health information need by formulating health search queries on Web search engines. In general, each health search query indicates some health information need. OHISs can express their information need either by formulating search queries using keywords or asking questions (wh-questions and yes/no questions). For this analysis, we considered the following wh-questions (lexicon): “What”, “How”, “?”, “When”, “Why”, and others (“Who”, “Where”, “Which”). Note that although “?” does not come under the wh-questions category, we have included it for simplicity. Yes/No questions are usually used to check factual information; for example, whether coffee is bad for the heart. In this analysis, we considered yes/no questions that start with “Can”, “Is”, “Does”, “Do”, “Are”, and others (“Could”, “Should”, “Will”, “Would”). Using the lexicon for wh-questions and yes/no questions, we performed text analysis on the search queries from PCs and SDs to count the number of queries with wh-questions and yes/no questions. Search queries that do not contain any question (wh- or yes/no) are classified as keyword-based. Additionally, for different wh- and yes/no questions, we computed their usage frequency in search queries from PCs and SDs.

Linguistic Analysis of Health Queries

Previous studies in generic search query log analysis have identified that understanding the linguistic structure, including phrase identification, entity spotting and descriptiveness (level of context), of search queries can improve Web Information Retrieval systems [Barr et al. 2008; Croft et al. 2010]. However, these efforts have not been applied extensively to health search queries, and hence in order

⁹<http://www.zyzyva.net/wordlists.shtml>

¹⁰<http://hunspell.sourceforge.net/>

to understand the linguistic structure of health queries, we performed part-of-speech analysis on search queries using Stanfords POS tagger [Toutanova et al. 2003]. For this analysis, we considered nouns, verbs, adjectives and adverbs. We mapped all subtypes in part-of-speech (eg, proper nouns, common nouns, compound nouns) to the main part-of-speech (eg, nouns). We analyzed the usage of different part-of-speech types in health queries based on their usage frequency in the PCs and SDs search queries.

6.3 Results

Top Health Queries

Most of the top search queries from both PCs and SDs are for symptom descriptions (eg, “lupus symptoms”). Another common way an OHIS searches for health information is by disease name (eg, “Lupus”). Chronic diseases (cancer, cardiovascular disease, diabetes) and diet (Mediterranean diet, gluten free food) are also searched often. Based on the top 100 search queries from PCs and SDs, we found that 48.49% of the search queries are different between PCs and SDs. However, due to the Mayo Clinic business confidentiality, we are not in a position to disclose the actual top search queries and numbers publicly.

Health Categories

While searching for health information, one in every three OHIS searches for “Symptoms” (Figure 6.2). Other popular health categories are “Causes” and “Treatments & Drugs”. Our analysis shows that the distribution of search queries for different health categories differs with the device used for the health search. At the same time, both PCs and SDs follow a similar pattern for distribution of the search queries between health categories. The percentage of OHIS searching for “Symptoms” is higher from SDs as compared to that from PCs. While for other health categories, the percentage of queries from PCs is slightly higher than that of SDs. Interestingly, one of the least searched health

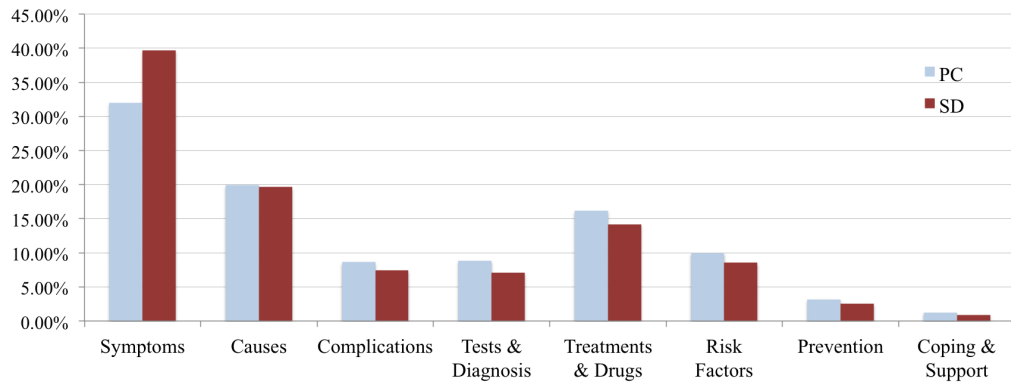


Figure 6.2: Distribution of the search queries by health categories.

		Personal computers		Smart device	
		QwoR %	QWR%	QwoR %	QWR %
Gender	Men	25.62	17.25	28.48	18.17
	Women	74.38	82.75	71.52	81.83
Age group	Children	66.55	59.60	79.33	74.39
	Teen	7.25	5.08	3.69	2.37
	Adults	18.60	31.68	13.64	21.72
	Elders	7.60	3.64	3.34	1.52
Temporal	Morning	26.85	29.93	31.93	39.14
	Afternoon/Evening	5.84	4.39	4.10	1.73
	Night	67.31	65.68	63.98	59.13

Table 6.1: Categorization of health search queries based on the information mentioned in the queries such as gender, age group, and temporal information (June 2011-May 2013)

categories is “Prevention”.

Categorization Based on the Information Mentioned in the Health Queries

The following are some of our observations based on the information referenced in the search queries (Table 6.3). The data indicate that the number of search queries mentioning words related to women's health is considerably higher compared to that of men. This implies that OHIS search for health information specifying women more often. The percentage of OHIS who use words related to “woman” in search queries is higher for PCs compared to SDs. Considering age group-related search queries, more than 60% of the queries are related to children. The percentage of OHIS that mention terms related to children in search queries is much higher for SDs compared to PCs. When considering a mention of the time of day in search queries, terms related to “Night” are mentioned most often (>60%) followed by words related to “Morning”. Very few search queries have words related to “Afternoon” and “Evening”. The percentage of OHIS using words related to “Morning” in search queries is higher for SDs compared to PCs, while the percentage of OHIS mentioning words related to “Night” in search queries is higher for PCs.

Health Query Length

The average search query length (Figures 6.3 and 6.4) for QwoR (PCs: 4.82 words and 26.73 characters; SDs: 5.33 words and 27.41 characters) is much larger than the average length of QwR (PCs: 2.90 words and 17.61 characters; SDs: 3.29 words and 18.86 characters). This indicates that longer search queries result in fewer repetitions, while shorter queries tend to be repeated more often. The analysis, although derived from a limited dataset, implies that in general health search queries tend to be longer than general search queries (not specific to one domain), as the average length of general search query from PCs is 2-2.35 words [Silverstein et al. 1999; Spink et al. 2001] and from SDs is 2.3 words [Kamvar and Baluja 2006]. This potentially indicates that OHISs describe their health information needs in more detail by adding relevant health context to the search query. Surprisingly, the average length of search query from SDs for both QwoR and QwR is slightly larger

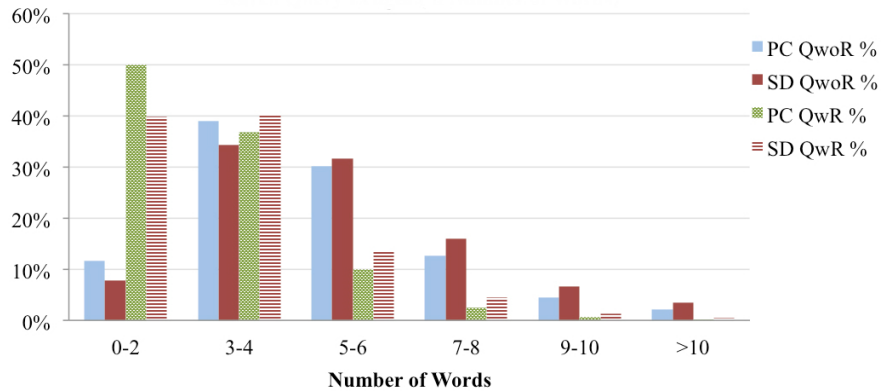


Figure 6.3: Distribution of the search queries by number of words and number

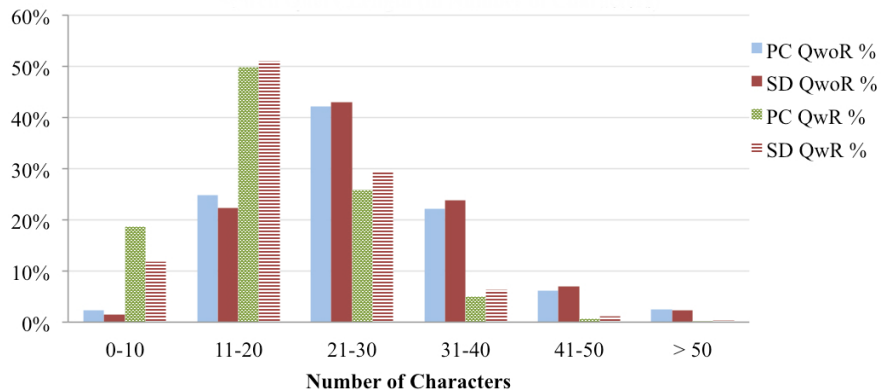


Figure 6.4: Distribution of the search queries by number of characters.

than queries from PCs.

Usage of Query Operators and Special Characters

In considering both PCs and SDs, approximately 10% of QwoR and 3% of QwR use at least one query operator. For QwR, the percentage of OHIS who use query operators in search queries is higher for SDs than PCs, while in the case of QWOR it is higher for PCs. AND is the most popular operator, followed by OR and “+”. Overall variations of “and” (AND, &, +) operators comprise more than 90% of operator usage. Considering QwoR, OHIS use AND OR query operators more often from SDs than that from PCs. Considering both PCs and SDs, around 10% of QwoR and

		Personal computers		Smart device	
		QwoR %	QwR %	QwoR %	QwR %
Number of operators	0	90.08	97.35	90.23	96.53
	>0	9.92	2.65	9.77	3.47
Query operators usage	AND	78.96	86.53	82.01	85.05
	+	11.24	4.37	6.29	3.08
	OR	6.95	5.2	8.74	6.78
	&	2.63	1.42	2.57	1.28
	Other	0.24	2.49	0.4	3.82
Special characters	0	89.02	95.66	90.54	96.72
	>0	10.98	4.34	9.46	3.29
Spelling mistakes	0	68.21	87.47	69.07	87.88
	>0	31.8	12.54	30.94	12.12

Table 6.2: Usage of query operators and special characters (June 11-May 13).

4% of QwR have at least one special character (Table 6.3). The percentage of OHIS using special characters in search queries is higher for PCs compared to SDs.

Misspellings in Health Queries

For QwoR and QwR, approximately 31% and 12% of queries, respectively, have at least one spelling mistake (Table 6.3). OHISs make slightly more spelling mistakes while searching health information from PCs than SDs.

Types of Health Queries

As indicated by the analysis in 6.5, OHISs predominantly formulate search queries using keywords, though wh-questions and yes/no questions are also substantial. Considering QwoR, OHISs ask

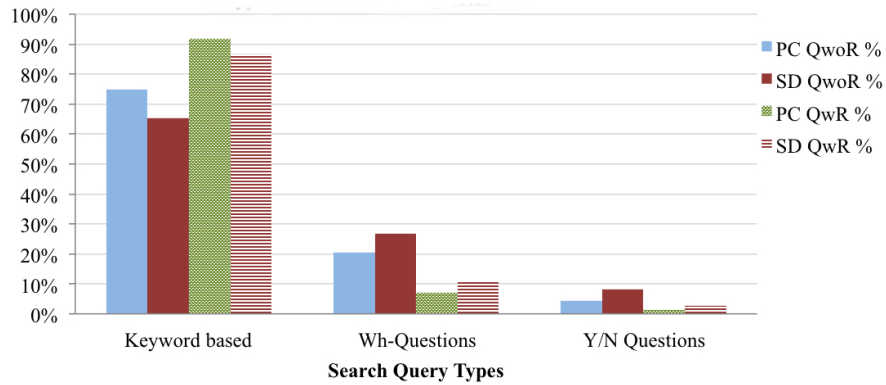


Figure 6.5: Types of health search queries (how health information need is expressed).

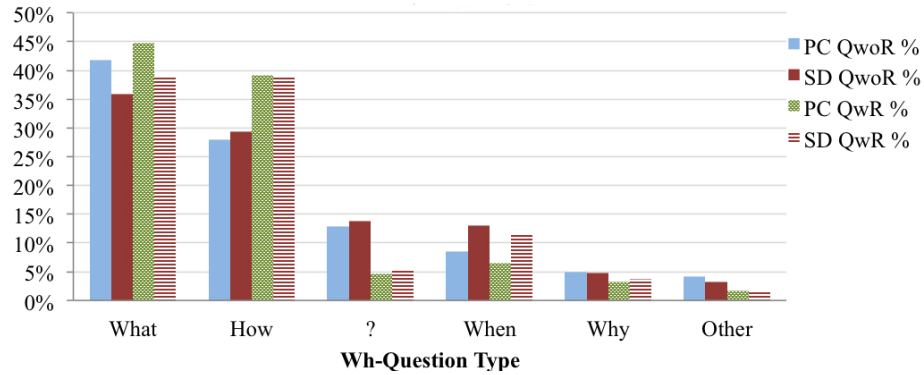


Figure 6.6: Distribution of the search queries based on type of wh-questions.

more (wh- and yes/no) questions from SDs than PCs. In wh-questions (6.6), OHISs mostly use “What” and “How” in the search queries, and both of them generally signify that more descriptive information is needed. OHISs ask more temporal questions (“When”) using SDs than PCs, while OHISs ask more “What” questions using PCs than SDs. In yes/no questions (6.7), OHISs generally start search queries with “Can”, “Is”, and “Does”. OHISs ask more yes/no questions starting with “Can” using SDs than using PCs, while the percentage of questions starting with “Is” and “Does” comes more from PCs.

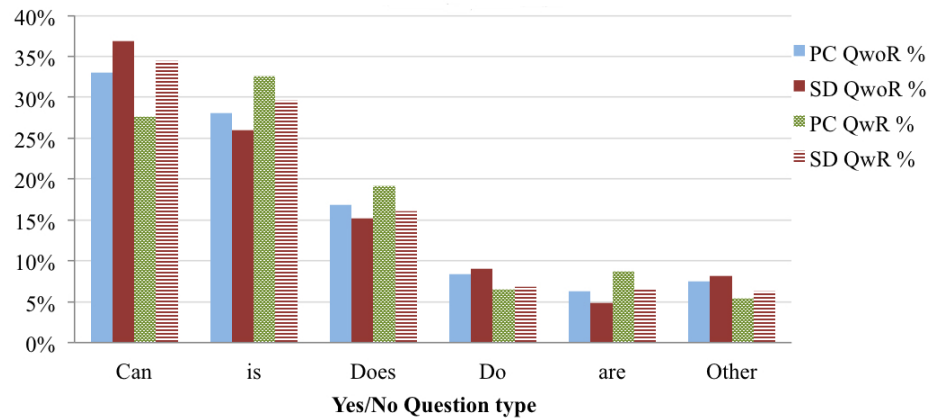


Figure 6.7: Distribution of the search queries based on type of yes/no questions.

Linguistic Analysis of Health Queries

In health search queries, nouns typically denote entities like disease names, health categories, etc. Almost all health search queries have at least one noun. In the case of QwR, most of the search queries (>70%) have 1-2 nouns, while in the case of QwoR, most of the search queries (>60%) have 2-3 nouns. There is no considerable difference in noun usage between PCs and SDs. A verb conveys an action or an occurrence, for example “how to control (verb) diabetes (noun)?”. Considering QwoR, OHIS use at least one verb in 37% of queries from PCs and 47% in queries from SDs. Adverbs are words that modify a verb, an adjective, and another adverb, while an adjective is a “describing” word, giving more information about the object signified; for example, “extremely (adverb) bad (adjective) stomach (noun) pain (noun)”. Very few search queries have at least one adverb. Considering QwoR, 45.66% of the queries from PCs and 48.50% of the queries from SDs have at least one adjective. This indicates that the percentage of search queries with at least one verb/adverb/adjective is higher for SDs than for PCs (see Table 6.3).

		Personal computers		Smart device	
		QwoR %	QwR %	QwoR %	QwR %
Noun	0	0.96	3.19	1.11	1.67
	1	14.31	28.17	14.52	26.93
	2	36.01	46.87	36.97	47.38
	3	31.34	17.75	31.61	19.79
	>3	17.37	4.01	15.8	4.23
verb	0	62.96	83.34	53.09	78.96
	>0	37.04	16.66	46.92	21.05
Adverb	0	93.86	95.56	91.01	95.38
	>0	6.15	4.45	9	4.62
Adjective	0	54.32	69.71	51.51	66.14
	>0	45.68	30.3	48.5	33.87

Table 6.3: Linguistic analysis of health search queries (June 2011-May 2013).

6.4 Discussion

6.4.1 Overview

Increasingly, individuals are actively participating in learning and managing their health by leveraging online resources. The percentage of people using the Internet and the usage of smart devices for health information searching is increasing rapidly. PCs and SDs have very distinct characteristics in terms of readability, user experience, accessibility, etc. These distinct characteristics provide some pros and cons for PCs and SDs: Web browsing and readability are better on PCs while accessibility is better for SDs. Also socioeconomic factors, such as age, gender, income level, education, familiarity with new technologies and devices [Fox and Duggan 2013; 2012], play an important role in the usage of PCs and SDs in general and for online health information seeking. Device characteristics and socioeconomic differences in device usage have an effect on OHISB [Fox and Duggan 2013; 2012; Higgins et al. 2011]. Therefore, in order to improve the health information searching process, it is necessary to understand both aspects, that is, how an OHIS searches for health information and how device choice influences online health information seeking.

In this study, we performed a comparative analysis on the most frequent health search queries launched from PCs and SDs to understand the effects of device type (PCs vs SDs) used for online health information seeking. The analysis dataset consists of search queries between June 2011 and May 2013, which were submitted from Web search engines and directed OHISs to the Mayo Clinic website. The website is visited by millions of unique OHIS every day, and it offers an identical appearance and accessibility for both PCs and SDs using standard Web search engines and Web browsers.

6.4.2 Principal Results

Following are some of the insights that surfaced from this study. Most of the top search queries from both PCs and SDs are related to symptoms, health conditions, chronic diseases, and diet. Our top search query analysis indicates that the device used has a significant effect on health information searching and the health information searched via different devices is also different (48.49%). While searching for health information, one in every three OHISs searches for “Symptoms”. Other popular health categories that OHISs search for are “Causes” and “Treatments & Drugs”. The analysis suggests that the distribution of search queries for different health categories differs with the device used for health search. Even though most of the diseases can be prevented with some lifestyle and diet changes, very few OHIS search for preventive health information. This highlights the fact that we need to promote preventive health care more vigorously.

While searching for health information, OHISs specify words related to women and children more often than that of men and any other age group. The higher percentage of women seeking online health information could be a reason [Fox and Duggan 2013; Higgins et al. 2011]. The percentage of OHISs who use words related to “women” and “night” in search queries is higher for PCs than for SDs, while “children” and “morning” are higher for SDs compared to PCs. Health search queries are longer than general search queries, which implies that OHISs describe health information need in more detail. Longer search queries also denote OHIS's interest in more specific information about the disease; subsequently, OHISs use more words to narrow down to a particular health topic. The average health search query length from SDs is longer than that of PCs, and while typing on SDs is slower and more difficult than typing on PCs, we posit that OHISs might be relying more on Web search engines' auto-completion functionality, as well as on most devices' speech recognition facilities, which might be increasing the length of search queries from SDs as compared to that from PCs. These results highlight the differences between usage of PCs and SDs for online health information seeking. The findings can be used by health websites and health application developers

to better understand OHISB for PCs and SDs, understand OHIS's health information needs, and better organize health information content for PCs and SDs users.

For PCs and SDs, 1 in 3 QwoR, and 1 in 10 QwR contained at least one spelling mistake. These mistakes place a burden on the search process and may lead users to incorrect or irrelevant information. The search engine's auto-completion feature, spelling correction/suggestion, and devices' speech recognition facilities might be contributing to reducing misspelled words in search queries. Almost all health search queries have at least one noun. In addition to nouns, OHISs use verbs, adverbs, and adjectives while formulating search queries to provide more context for the topic of interest. The percentage of search queries with at least one verb/adverb/adjective is higher for SDs as compared to PCs. This implies that health search queries from SDs are more descriptive as compared to queries from PCs. OHISs formulate search queries by using keywords most frequently, followed by wh-questions and yes/no questions. Considering QwoR, OHIS ask more questions via SD than PC. In wh-questions, OHISs mostly use "What" and "How" in search queries, and both of them generally signify a need for more descriptive information while search queries in the form of yes/no questions indicate interest in factual information.

Since search queries are a fundamental part of health information searching, it is essential that we understand characteristics of health search queries and the role of the device used for searching. This study provides useful insights for online health information retrieval systems. The linguistic structure of a search query has implications in information retrieval using Web search engines [Barr et al. 2008; Croft et al. 2010]. Cory Barr et al [Barr et al. 2008][42] highlight the importance of recognizing part-of-speech information of the input search query to improve search results and demonstrate that the part-of-speech is a significant feature for information retrieval. Our study provides distribution of part-of-speech in health search queries from PCs and SDs. Expressiveness or descriptiveness of the search queries has a significant impact on quality of the search results using Web search engines [Croft et al. 2010]. Phan et al [Phan et al. 2007] specify that with the increase in search query length, the descriptiveness of the query increases. Our study gives basic understanding

about health search query descriptiveness based on health query length and part-of-speech analysis. Previous research in information retrieval have identified various important features of search queries such as usage of search query operators [Eastman and Jansen 2003][58], misspellings, query length [Phan et al. 2007], query type (keyword-based, wh-questions, yes/no questions), and part-of-speech [Barr et al. 2008; Croft et al. 2010]. We presented a comprehensive analysis of these features for health search queries via PCs and SDs.

6.4.3 Comparison With Related Work

This study contributes a comparative analysis performed on large-scale health search queries to understand the effects of device type (SDs vs PCs) used on OHISB. As discussed in the “Background and Significance” section, previous efforts have used several approaches to understand OHISB including (1) focus groups and user surveys, and (2) analyzing health-related Web search query logs. To the best of our knowledge, there is not much research on understanding the effect of devices on online health search behavior. In our work, we bridge this knowledge gap by analyzing more than 100 million health search queries from PCs and SDs to understand how device choice influences online health information seeking. In addition, we presented analysis for both QwR and QwoR in order to avoid bias from queries with a high number of repetitions. Moreover, we analyzed linguistic structure of health search queries from PCs and SDs, which has implications for Web search engines and information retrieval systems.

6.4.4 Limitations

The results of this study are derived from analysis limited to health search queries from Web search engines that led users to Mayo Clinic website. Even though Mayo Clinic web pages often ranked high in Web search engines, not all health information seekers visited the Mayo Clinic website. Also, this analysis is based on the top one million health queries per month (PCs/SDs) rather than the entire health traffic to Mayo Clinic site. In this work, we considered search queries from smartphones and

tablets into same categories (ie, smart devices) as the search queries are differentiated based on the operating system of the device used for search, and not the type of specific device per se (eg, Apple iPhone vs iPad vs Android phone). The focus of this study is limited to analysis of a search query log, and we have not analyzed associated socioeconomic factors due to anonymized nature of the data. Previous studies have identified that socioeconomic factors such as age, gender, education, and income have an effect on device usage and OHISB. Further research in analyzing health search queries based on socioeconomic factors can extend our knowledge about how socioeconomic factors affect health search query formation and the type of health information searched.

6.4.5 Future Work

In the future, we will extend this work by performing a semantic analysis on the data using biomedical knowledge bases and ontologies. Specifically, we plan to leverage insights from this work and use semantic Web technologies to facilitate health search experience by developing more advanced next-generation knowledge and content delivery systems.

6.5 Conclusion

We presented a comprehensive analysis of large-scale health search queries from personal computers (desktops/laptops) and smart devices (smartphones/tablets) in order to understand the effects of device type on the features that contribute in the expression of search intent. We noted that online health information search behavior differs from general online information search. Also, the type of device used for online health information search plays an important role and alters the expression of search intents. A greater understanding of OHIS's needs, especially how they search and what they search for, may help us understand behavioral changes that will lead to improvement in online health information seeking and a more balanced approach to wellness and prevention. This study extends our knowledge about online health information search behavior, difference in the expression

of search intents by device types and provides useful information for Web search engines, health-centric websites, health care providers, and health carecentric application developers. Finally, we anticipate that this work will help empower OHISs in their quest for health information and facilitate their health information search efforts by enabling the development of more advanced next-generation knowledge and content delivery systems.

7

Conclusions

7.1 Summary

Search intent mining can help Web search engines to enhance their ranking of search results, enabling new search features like personalization, search result diversification, and the recommendation of relevant ads. By understanding the domain of a search query, a search engine can return more relevant and essential results, complimentary structured information, and targeted ads rather than providing keyword-based results. While state-of-the-art techniques can identify the domain of the queries (e.g. sports, movies, health), identifying domain-specific intent is still an open problem. Among all the topics available on the Internet, health is one of the most important in terms of impact on the user and forms one of the most frequently searched areas. In this dissertation, we presented a knowledge-driven approach for domain-specific search intent mining with a focus on health-related search queries.

First, we identified 14 consumer-oriented health search intent classes based on inputs from focus group studies and based on analyses of popular health websites, literature surveys, and an empirical study of search queries. We defined the problem of classifying millions of health search queries into zero or more intent classes as a multi-label classification problem. Popular machine learning

approaches for multi-label classification tasks (namely, problem transformation and algorithm adaptation methods) were not feasible as manually annotating search queries with multiple intent classes is very labor intensive and slow. Furthermore, classifiers trained for one disease may not work for other diseases as the symptoms, treatments, drugs, and medications vary by the disease. At the same time, there are several biomedical knowledge sources that encode vast clinical knowledge in a structured way that can be easily shared and reused by both humans and computers. They contain many millions of individual entities, their mappings into semantic classes, and the relationships between entities. We leveraged this rich biomedical knowledge to address search intent mining problem in a disease agnostic manner.

We used 10 million cardiovascular diseases-related search queries from Mayo Clinic to conduct our experiments and macro-average precision recall as our evaluation metrics. First, we have utilized Semantic Types associated with the intent classes as a baseline approach to classify search queries into intent classes (precision: 0.5432 , recall: 0.6203 ,and F1-score: 0.5791). We iteratively improved the baseline approach using a) semantic concepts, b) excluding misclassified Semantic Types and Semantic Concepts, c) addressing concept identification challenges by incorporating advanced text analytics techniques such as word sense disambiguation and maximal phrase detection, and d) using Consumer Health Vocabulary (CHV) from UMLS which maps consumer oriented terms to the associated medical terms.

While CHV from UMLS is very useful, it is manually curated and has limited coverage. This vocabulary gap is a major challenge for the health search intent mining problem since a large of portion of health search queries are submitted by laymen. We leveraged crowd-sourced knowledge from Wikipedia to improve the coverage of the CHV. We developed a pattern-based information extractor that extracts candidate pairs of CHV terms and medical terms from medical health-related Wikipedia pages. We used a hypothesis-based approach to identify CHV terms. As compared to most CHV generation approaches, our approach is automated and does not require manual review of CHV terms from domain experts. Furthermore, it uses knowledge from Wikipedia that is being

continuously updated with emerging health terms.

Finally, we developed a semantics-driven, rule-based intent mining approach by leveraging rich background knowledge encoded in UMLS and Wikipedia. Based on the evaluation, our classification approach had very good precision: 0.8842, recall: 0.8642, and F1-Score: 0.8723. Most of the 10 million queries are classified into either one (47.72%) or two intent(39.87%) classes. This approach can identify search intent in a disease-agnostic manner and has been evaluated on the three major chronic diseases: cardiovascular diseases, diabetes, and cancer. We selected chronic diseases for the experiments and evaluations due to their very high prevalence and the fact that they are by far the leading causes of mortality in the world.

Next, we applied the search intent mining algorithm on health related Twitter data. While users often turn to search engines to learn about health conditions, a surprising amount of health information is also shared and consumed via social media, such as the public social platform Twitter. In some cases, people prefer Twitter as an information source as compared to the traditional information sources due to its information aggregation capabilities. Although Twitter is an excellent information source and has many advantages, identification of informative tweets from the deluge of tweets is a major challenge. Furthermore, the informativeness of a tweet is very subjective. To facilitate identification of informative and trustworthy content in tweets, it is crucial to develop an effective classification system that can objectively classify tweets as informative or uninformative. Thus, in order to address this problem, we have abstracted out the subjective nature of the informativeness problem and objectively studied what features contribute to informativeness. To this end, we developed a hybrid approach using rule-based filtering and supervised classification for the identification of the informative tweets.

In the rule-based filtering step, we used the following filters: tweets in the English language, tweets with URLs, minimum tweet length and minimum Google PageRank of 5 for URLs. We also filtered-out duplicate tweets, broken and non-functional URLs. Using the rule-based filtering, we reduced the experimental dataset from 40K tweets to 6.3K tweets (84.25% reduction in the dataset).

For the supervised classification, we selected a Naive Bayes classifier as it was very fast (a crucial factor for classifying millions of tweets in a timely manner) and had competitive performance with respect to the other classifiers. For the classification, we used the following features associated with the tweets and their URLs: 1) text features: n-grams, length of the tweet, number of special characters, and POS tags; 2) author features: number of follow-followers, number of tweets, and the authors' credibility or influence, i.e. Klout score; 3) popularity features: Twitter (shares, retweets), Facebook (shares, likes, and comments), and Google Plus shares; and 5) Google PageRank of the URLs in the tweets. Using a Naive Bayes classifier and above mentioned features, we classified 80.93% (precision) of the tweets correctly.

We also presented a system, Social Health Signals, which aggregates the informative health information shared on the Twitter platform in near real time. To enable efficient browsing of the health information on Social Health Signals, we are using our search intent mining algorithm which classifies informative tweets and health news into consumer-oriented health categories like symptoms, living with, food and diet, prevention and treatments. Finally, we presented a comprehensive analysis of large-scale health search queries from personal computers (desktops/laptops) and smart devices (smartphones/tablets) in order to understand the effects of device type on the expression of search intents. We concluded that online health information search behavior differs from general online information search. Also, the type of device used for online health information search plays an important role and alters the expression of search intents.

7.2 Future Directions

There are several directions that are worth exploring in the future.

For the past 10 hours I've been experiencing a semi sharp pain in my upper right chest just below my armpit. This pain appears anywhere from every two and a half minutes to ten or fifteen minutes. I also have some stomach ache and dry mouth. I monitor my blood pressure is averages 130/90 with a average heart rate of 80. My cardiologist has been treating me since 1 year for high colesterol, gout and hypertension with great success. Also I have diabetes and I am taking Metformin and mevacor. I have an appointment with my cardiologist after 2 weeks. However I am wondering should I go to ER? BTW I am 69 years old male.

Source: DailyStrength forum

Figure 7.1: Medical question posted by a layman on one medical question-answering forum (DailyStrength)

Medical Question Answering

With initiatives like IBM Watson Health, we are moving towards cognitive assistants that can help healthcare providers in clinical decision making. Medical Question Answering (QA) is one of the prominent areas where these cognitive assistants can help healthcare providers. In medical QA systems, it is crucial to understand consumers' questions, extract topics from the questions, and direct it to healthcare professionals who are experts in the extracted topics. The search intent mining algorithm presented in this dissertation can be extended for this task. In a nutshell, the techniques used in search intent mining problem can be leveraged to automatically extract structured medical information from unstructured (free text) medical questions submitted by laymen. Let us consider the following use-case scenario (Figure 7.1) consisting of a medical question posted by a layman on one medical QA forum (DailyStrength).

As shown in (Figure 7.2 and 7.3), using the techniques developed for search intent mining algo-

For the past 10 hours I've been experiencing a semi sharp pain in my upper right chest just below my armpit. This pain appears anywhere from every two and a half minutes to ten or fifteen minutes. I also have some stomach ache and dry mouth. I monitor my blood pressure is averages 130/90 with a average heart rate of 80. My cardiologist has been treating me since 1 year for high colesterol, gout and hypertension with great success. Also I have diabetes and I am taking Metformin and mevacor. I have an appointment with my cardiologist after 2 weeks. However I am wondering should I go to ER? BTW I am 69 years old male.

Figure 7.2: Information extraction using search intent mining algorithm

Misspellings	experiencing => experiencing colesterol! => cholesterol
Consumer Health Vocabulary	dry mouth => Xerostomia
Symptom	chest pain stomach ache Xerostomia (dry mouth)
Diseases and Conditions	Gout Hypertension Diabetes
Drugs and Medication	Metformin Mevacor
Vital Signs	Blood pressure: 130/90 Heart rate: 80
Demographic Information	Age: 69 Gender: Male

Figure 7.3: Structured medical information extracted from unstructured medical question using techniques used in search intent mining algorithm.

rithm, we can process the unstructured medical text and extract useful medical information. Such information can be further used in an automated or semi-automated manner to answer the medical questions. In a semi-automated approach the extracted information is used 1) to direct questions to the appropriate healthcare provider and 2) to get a structured summary of the medical question for the healthcare provider that can save their valuable time needed for manually going through the medical question and noting down symptoms, medications, etc. In an automated approach for question-answering, the extracted information can be used to generate set hypothesis and validate them based on the clinical knowledge. IBM Watson Health is working on automated question answering and the Watson team has acknowledged that the structured information extracted by the algorithms presented in this dissertation can provide essential information for automated medical question answering.

Health Information Intervention

Health information or any information is useful for a reader only if the information is relevant to him. Health information intervention can be very beneficial for a patient if he can learn about medical conditions, symptoms, and treatment options that he may need to know about and would not think to check the Internet on his own. Such information can be valuable, relevant, and even lifesaving for patients. In order to do targeted information intervention, it is crucial to identify users' interests. Rather than relying on the unrealistic assumption that people will precisely specify their interest. We can identify the users' interests from their health search queries. Users' health information interests can be short-term (e.g. seasonal diseases, curiosity for a health condition) and/or long-term (e.g. chronic diseases, interest in healthy lifestyle). We can create user interest profiles based on both their short and long-term search histories for personalized health information interventions.

Search Personalization

Every user has a distinct background and a specific goal when searching for information on the Web. The goal of Web search personalization is to tailor search results to a particular user based on that user's interests and preferences. For constructing the necessary user interest profiles for search personalization, evidence of a user's interests can be mined from observed past behaviors. A user's history of queries provides cues to construct the user's interest profile. We can extend the search intent mining approach presented in the dissertation to build user profiles that can represent the health-related topical interests of the users. We can model users' interest profiles from different temporal views of their history of interaction with the search engine. User profiles can be created by classifying the terms from users' queries into search intent classes. Such user profiles can be useful for personalized ranking of the search results, i.e. better search result relevance and targeted advertisements.

References

- AHMAD, F., HUDAK, P. L., BERCOVITZ, K., HOLLENBERG, E., AND LEVINSON, W. 2006. Are physicians ready for patients with internet-based health information? *Journal of Medical Internet Research* 8, 3, e22.
- ALLEN, B. 1991. Topic knowledge and online catalog search formulation. *The Library Quarterly*, 188–213.
- ANDREASSEN, H. K., BUJNOWSKA-FEDAK, M. M., CHRONAKI, C. E., DUMITRU, R. C., PUDULE, I., SANTANA, S., VOSS, H., AND WYNN, R. 2007. European citizens' use of e-health services: a study of seven countries. *BMC public health* 7, 1, 1.
- ARAZY, O. AND NOV, O. 2010. Determinants of wikipedia quality: the roles of global and local contribution inequality. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 233–236.
- ARGUELLO, J., DIAZ, F., CALLAN, J., AND CRESPO, J.-F. 2009. Sources of evidence for vertical selection. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 315–322.
- ARGUELLO, J., DIAZ, F., AND PAIEMENT, J.-F. 2010. Vertical selection in the presence of unlabeled verticals. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 691–698.

- ARONSON, A. R. AND LANG, F.-M. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17, 3, 229–236.
- ASHKAN, A., CLARKE, C. L., AGICHTEN, E., AND GUO, Q. 2009. Classifying and characterizing query intent. In *Advances in Information Retrieval*. Springer, 578–586.
- ATKINSON, N., SAPERSTEIN, S., AND PLEIS, J. 2009. Using the internet for health-related activities: findings from a national probability sample. *Journal of medical Internet research* 11, 1, e5.
- AUER, S., BIZER, C., KOBILAROV, G., LEHMANN, J., CYGANIAK, R., AND IVES, Z. 2007. *Dbpedia: A nucleus for a web of open data*. Springer.
- AYERS, S. L. AND KRONENFELD, J. J. 2007. Chronic illness and health-seeking information on the internet. *Health: 11, 3, 327–347*.
- BAEZA-YATES, R., CALDERÓN-BENAVIDES, L., AND GONZÁLEZ-CARO, C. 2006. The intention behind web queries. In *String processing and information retrieval*. Springer, 98–109.
- BAKER, L. D. AND MCCALLUM, A. K. 1998. Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 96–103.
- BARR, C., JONES, R., AND REGELSON, M. 2008. The linguistic structure of english web-search queries. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1021–1030.
- BEEFERMAN, D. AND BERGER, A. 2000. Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 407–416.
- BEITZEL, S. M., JENSEN, E. C., CHOWDHURY, A., AND FRIEDER, O. 2007. Varying approaches to topical web query classification. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 783–784.

- BEITZEL, S. M., JENSEN, E. C., FRIEDER, O., GROSSMAN, D., LEWIS, D. D., CHOWDHURY, A., AND KOLCZ, A. 2005. Automatic web query classification using labeled and unlabeled training data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 581–582.
- BEITZEL, S. M., JENSEN, E. C., FRIEDER, O., LEWIS, D. D., CHOWDHURY, A., AND KOLCZ, A. 2005. Improving automatic query classification via semi-supervised learning. In *Data Mining, Fifth IEEE international Conference on*. IEEE, 8–pp.
- BENIGERI, M. AND PLUYE, P. 2003. Shortcomings of health information on the internet. *Health promotion international* 18, 4, 381–386.
- BERLAND, G. K., ELLIOTT, M. N., MORALES, L. S., ALGAZY, J. I., KRAVITZ, R. L., BRODER, M. S., KANOUSE, D. E., MUÑOZ, J. A., PUYOL, J.-A., LARA, M., ET AL. 2001. Health information on the internet: accessibility, quality, and readability in english and spanish. *Jama* 285, 20, 2612–2621.
- BESSELL, T. L., SILAGY, C. A., ANDERSON, J. N., HILLER, J. E., AND SANSOM, L. N. 2002. Measuring prevalence: Prevalence of south australia’s online health seekers. *Australian and New Zealand journal of public health* 26, 2, 170–173.
- BI, W. AND KWOK, J. T. 2011. Multi-label classification on tree-and dag-structured hierarchies. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 17–24.
- BLOCK, L., HABICHT, R., WU, A. W., DESAI, S. V., WANG, K., SILVA, K. N., NIESSEN, T., OLIVER, N., AND FELDMAN, L. 2013. In the wake of the 2003 and 2011 duty hours regulations, how do internal medicine interns spend their time? *Journal of general internal medicine* 28, 8, 1042–1047.
- BODIE, G. D. AND DUTTA, M. J. 2008. Understanding health literacy for strategic health marketing:

- ehealth literacy, health disparities, and the digital divide. *Health Marketing Quarterly* 25, 1-2, 175–203.
- BONNAR-KIDD, K. K., BLACK, D. R., MATTSON, M., AND COSTER, D. 2009. Online physical activity information: will typical users find quality information? *Health communication* 24, 2, 165–175.
- BOYD, D., GOLDER, S., AND LOTAN, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 1–10.
- BRENNAN, P. F. AND ARONSON, A. R. 2003. Towards linking patients and clinical information: detecting umls concepts in e-mail. *Journal of biomedical informatics* 36, 4, 334–341.
- BRINKER, K. AND HÜLLERMEIER, E. 2007. Case-based multilabel ranking. In *IJCAI*. 702–707.
- BRODER, A. 2002. A taxonomy of web search. In *ACM Sigir forum*. Vol. 36. ACM, 3–10.
- BRODER, A. Z., FONTOURA, M., GABRILOVICH, E., JOSHI, A., JOSIFOVSKI, V., AND ZHANG, T. 2007. Robust classification of rare queries using web knowledge. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 231–238.
- BROKOWSKI, L. AND SHEEHAN, A. H. 2009. Evaluation of pharmacist use and perception of wikipedia as a drug information resource. *Annals of Pharmacotherapy* 43, 11, 1912–1913.
- BROWNSTEIN, J. S., FREIFELD, C. C., AND MADOFF, L. C. 2009. Digital disease detection harnessing the web for public health surveillance. *New England Journal of Medicine* 360, 21, 2153–2157.
- CAO, H., HU, D. H., SHEN, D., JIANG, D., SUN, J.-T., CHEN, E., AND YANG, Q. 2009. Context-aware query classification. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 3–10.

- CAO, H., JIANG, D., PEI, J., HE, Q., LIAO, Z., CHEN, E., AND LI, H. 2008. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 875–883.
- CARNEIRO, H. A. AND MYLONAKIS, E. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases* 49, 10, 1557–1564.
- CELIKIYLMAZ, A., HAKKANI-TÜR, D., AND TÜR, G. 2011. Leveraging web query logs to learn user intent via bayesian discrete latent variable model. In *Proceedings of ICML*.
- CHAN, C. V. AND KAUFMAN, D. R. 2011. A framework for characterizing ehealth literacy demands and barriers. *Journal of Medical Internet Research* 13, 4, e94.
- CHAPMAN, K., ABRAHAM, C., JENKINS, V., AND FALLOWFIELD, L. 2003. Lay understanding of terms used in cancer consultations. *Psycho-Oncology* 12, 6, 557–566.
- CHAPMAN, W. W., BRIDEWELL, W., HANBURY, P., COOPER, G. F., AND BUCHANAN, B. G. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics* 34, 5, 301–310.
- CHAUDHURI, M. S., LE, M. T., WHITE, M. C., THOMPSON, H., AND DEMIRIS, G. 2013. Examining health information-seeking behaviors of older adults. *Computers, informatics, nursing: CIN* 31, 11, 547.
- CHEN, P. 2013. For new doctors, 8 minutes per patient. *New York Times*.
- CHENG, W., HÜLLERMEIER, E., AND DEMBCZYNSKI, K. J. 2010. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 279–286.
- CHERMAN, E. A., MONARD, M. C., AND METZ, J. 2011. Multi-label problem transformation methods: a case study. *CLEI Electronic Journal* 14, 1, 4–4.

- CHOI, N. G. AND DINITTO, D. M. 2013. The digital divide among low-income homebound older adults: Internet use patterns, ehealth literacy, and attitudes toward computer/internet use. *Journal of medical Internet research* 15, 5, e93.
- CLARE, A. AND KING, R. D. 2001. Knowledge discovery in multi-label phenotype data. In *Principles of data mining and knowledge discovery*. Springer, 42–53.
- CLARKE, C. L., CRASWELL, N., AND SOBOROFF, I. 2009. Overview of the trec 2009 web track. Tech. rep., DTIC Document.
- CONNOLLY, K. K. AND CROSBY, M. E. 2014. Examining e-health literacy and the digital divide in an underserved population in hawai'i. *Hawai'i journal of medicine & public health: a journal of Asia Pacific Medicine & Public Health* 73, 2, 44–48.
- CORBIN, J. AND STRAUSS, A. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.
- CRASWELL, N. AND SZUMMER, M. 2007. Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 239–246.
- CROFT, W. B., METZLER, D., AND STROHMAN, T. 2010. *Search engines: Information retrieval in practice*. Vol. 283. Addison-Wesley Reading.
- CUTILLI, C. C. 2010. Seeking health information: what sources do your patients use? *Orthopaedic Nursing* 29, 3, 214–219.
- DAI, M., SHAH, N., XUAN, W., ET AL. 2008. An efficient solution for mapping free text to ontology terms. amia summit on translational bioinformatics. *San Francisco CA*.
- DANG, V., XUE, X., AND CROFT, W. B. 2011. Inferring query aspects from reformulations using clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2117–2120.

- DE CHOUDHURY, M., MORRIS, M. R., AND WHITE, R. W. 2014. Seeking and sharing health information online: Comparing search engines and social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1365–1376.
- DE COMITÉ, F., GILLERON, R., AND TOMMASI, M. 2003. Learning multi-label alternating decision trees from texts and data. In *Machine Learning and Data Mining in Pattern Recognition*. Springer, 35–49.
- DENECKE, K. AND NEJDL, W. 2009. How valuable is medical social media data? content analysis of the medical web. *Information Sciences* 179, 12, 1870–1880.
- DOGAN, R. I., MURRAY, G. C., NÉVÉOL, A., AND LU, Z. 2009. Understanding pubmed® user search behavior through log analysis. *Database 2009*, bap018.
- DOING-HARRIS, K. M. AND ZENG-TREITLER, Q. 2011. Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of medical Internet research* 13, 2, e37.
- DRENTEA, P., GOLDNER, M., COTTEN, S., AND HALE, T. 2008. The association among gender, computer use and online health searching, and mental health. *Information, Communication & Society* 11, 4, 509–525.
- DUAN, Y., JIANG, L., QIN, T., ZHOU, M., AND SHUM, H.-Y. 2010. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 295–303.
- DUGGAN, M. AND SMITH, A. 2013. Cell internet use 2013. *Washington, DC: PewResearchCenter*.
- EASTMAN, C. M. AND JANSEN, B. J. 2003. Coverage, relevance, and ranking: The impact of query operators on web search engine results. *ACM Transactions on Information Systems (TOIS)* 21, 4, 383–411.

- ELISSEEFF, A. AND WESTON, J. 2001. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*. 681–687.
- ERDELEZ, S. 1997. Information encountering: a conceptual framework for accidental information discovery. In *Proceedings of an international conference on Information seeking in context*. Taylor Graham Publishing, 412–421.
- ERDELEZ, S. AND RIOUX, K. 2000. Sharing information encountered for others on the web. *The new review of information behaviour research* 1, January, 219–233.
- EYSENBACH, G. AND KÖHLER, C. 2002. How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. *Bmj* 324, 7337, 573–577.
- EYSENBACH, G. AND KÖHLER, C. 2004. Health-related searches on the internet. *Jama* 291, 24, 2946–2946.
- FLYNN, K. E., SMITH, M. A., AND FREESE, J. 2006. When do older adults turn to the internet for health information? findings from the wisconsin longitudinal study. *Journal of General Internal Medicine* 21, 12, 1295–1301.
- FOX, S. 2014. Pew internet & american life project report. 2013. *Pew Internet: Health* URL: <http://www.pewinternet.org/fact-sheets/health-fact-sheet/>[accessed 2014-02-21][WebCite Cache].
- FOX, S. AND DUGGAN, M. 2012. Mobile health 2012. pew internet and american life project.
- FOX, S. AND DUGGAN, M. 2013. Health online 2013. *Health*, 1–55.
- FOX, S. AND JONES, S. 2009. The social life of health information. pew internet.
- FOX, S. AND JONES, S. 2012. The social life of health information. pew internet & american life project 2009.

- FUJITA, S., MACHINAGA, K., AND DUPRET, G. 2010. Click-graph modeling for facet attribute estimation of web search queries. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 190–197.
- FÜRNKRANZ, J., HÜLLERMEIER, E., MENCÍA, E. L., AND BRINKER, K. 2008. Multilabel classification via calibrated label ranking. *Machine learning* 73, 2, 133–153.
- GAN, Q., ATTENBERG, J., MARKOWETZ, A., AND SUEL, T. 2008. Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web*. ACM, 49–56.
- GIACOMINI, M. K., COOK, D. J., GROUP, E.-B. M. W., ET AL. 2000. Users' guides to the medical literature: Xxiii. qualitative research in health care a. are the results of the study valid? *Jama* 284, 3, 357–362.
- GILES, J. 2005. Internet encyclopaedias go head to head. *Nature* 438, 7070, 900–901.
- GINSBERG, J., MOHEBBI, M. H., PATEL, R. S., BRAMMER, L., SMOLINSKI, M. S., AND BRILLIANT, L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232, 1012–1014.
- GODBOLE, S. AND SARAWAGI, S. 2004. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining*. Springer, 22–30.
- GUTIERREZ, N., KINDRATT, T. B., PAGELS, P., FOSTER, B., AND GIMPEL, N. E. 2014. Health literacy, health information seeking behaviors and internet use among patients attending a private and public clinic in the same geographic area. *Journal of community health* 39, 1, 83–89.
- HAMANN, J., MENDEL, R., BÜHNER, M., KISSLING, W., COHEN, R., KNIPFER, E., AND ECKSTEIN, H.-H. 2012. How should patients behave to facilitate shared decision making—the doctors view. *Health Expectations* 15, 4, 360–366.

- HAMMOND, B., SHETH, A., AND KOCHUT, K. 2002. Semantic enhancement engine: A modular document enhancement platform for semantic applications over heterogeneous content. *Real World Semantic Web Applications 92*, 29.
- HARIHARAN, B., ZELNIK-MANOR, L., VARMA, M., AND VISHWANATHAN, S. 2010. Large scale max-margin multi-label classification with priors. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 423–430.
- HARVEY, M., CRESTANI, F., AND CARMAN, M. J. 2013. Building user profiles from topic models for personalised search. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2309–2314.
- HERSKOVIC, J. R., TANAKA, L. Y., HERSH, W., AND BERNSTAM, E. V. 2007. A day in the life of pubmed: analysis of a typical day’s query log. *Journal of the American Medical Informatics Association 14*, 2, 212–220.
- HIGGINS, O., SIXSMITH, J., BARRY, M. M., AND DOMEGAN, C. 2011. A literature review on health information seeking behaviour on the web: a health consumer and health professional perspective.
- HOFFART, J., PREDA, N., SUCHANEK, F. M., AND WEIKUM, G. 2015. Knowledge bases for web content analytics. In *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1535–1535.
- HU, J., WANG, G., LOCHOVSKY, F., SUN, J.-T., AND CHEN, Z. 2009. Understanding user’s query intent with wikipedia. In *Proceedings of the 18th international conference on World wide web*. ACM, 471–480.
- HU, Y., QIAN, Y., LI, H., JIANG, D., PEI, J., AND ZHENG, Q. 2012. Mining query subtopics from search log data. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 305–314.

- HUMPHREY, S. M., ROGERS, W. J., KILICOGU, H., DEMNER-FUSHMAN, D., AND RINDFLESCH, T. C. 2006. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology* 57, 1, 96–113.
- HURFORD, J. R., STUDDERT-KENNEDY, M., AND KNIGHT, C. 1998. *Approaches to the evolution of language: Social and cognitive bases*. Cambridge University Press.
- JADHAV, A., ANDREWS, D., FIKSDAL, A., KUMBAMU, A., MCCORMICK, J. B., MISITANO, A., NELSEN, L., RYU, E., SHETH, A., WU, S., ET AL. 2014. Comparative analysis of online health queries originating from personal computers and smart devices on a consumer health information portal. *Journal of medical Internet research* 16, 7, e160.
- JADHAV, A., WU, S., SHETH, A., AND PATHAK, J. 2014a. Online information seeking for cardiovascular diseases: A case study from mayo clinic. In *MIE, Studies in health technology and informatics*. Vol. 2014. European Medical Informatics, 702.
- JADHAV, A. S. AND PATHAK, J. 2014. Comparative analysis of online health information search by device type.
- JADHAV, A. S., PUROHIT, H., KAPANIPATHI, P., ANANTHARAM, P., RANABAHU, A. H., NGUYEN, V., MENDES, P. N., SMITH, A. G., COONEY, M., AND SHETH, A. P. 2010. Twitris 2.0: Semantically empowered system for understanding perceptions from social data.
- JADHAV, A. S., SHETH, A. P., AND PATHAK, J. 2014. An analysis of mayo clinic search query logs for cardiovascular diseases.
- JADHAV, A. S., SONI, S., AND SHETH, A. P. 2015. Social health signals.
- JADHAV, A. S., WANG, W., MUTHARAJU, R., ANANTHARAM, P., NGUYEN, V., SHETH, A. P., GOMADAM, K., NAGARAJAN, M., AND RANABAHU, A. H. 2013. Twitris: socially influenced browsing.

- JADHAV, A. S., WU, S., SHETH, A. P., AND PATHAK, J. 2014b. What information about cardiovascular diseases do people search online?
- JANSEN, B. J. AND BOOTH, D. 2010. Classifying web queries by topic and user intent. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 4285–4290.
- JANSEN, B. J., BOOTH, D. L., AND SPINK, A. 2008. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management* 44, 3, 1251–1266.
- JANSEN, B. J., SPINK, A., BATEMAN, J., AND SARACEVIC, T. 1998. Real life information retrieval: A study of user queries on the web. In *ACM SIGIR Forum*. Vol. 32. ACM, 5–17.
- JOACHIMS, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 133–142.
- JONES, R. AND KLINKNER, K. L. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 699–708.
- JONQUET, C., SHAH, N., AND MUSEN, M. 2009. The open biomedical annotator. In *AMIA summit on translational bioinformatics*. 56–60.
- KAMVAR, M. AND BALUJA, S. 2006. A large scale study of wireless search behavior: Google mobile search. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 701–709.
- KESELMAN, A., SMITH, C. A., DIVITA, G., KIM, H., BROWNE, A. C., LEROY, G., AND ZENG-TREITLER, Q. 2008. Consumer health concepts that do not map to the umls: where do they fit? *Journal of the American Medical Informatics Association* 15, 4, 496–505.

- KITTUR, A., SUH, B., PENDLETON, B. A., AND CHI, E. H. 2007. He says, she says: conflict and coordination in wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 453–462.
- KIVITS, J. 2006. Informed patients and the internet a mediated context for consultations with health professionals. *Journal of health psychology* 11, 2, 269–282.
- KNAPP, C., MADDEN, V., MARCU, M., WANG, H., CURTIS, C., SLOYER, P., AND SHENKMAN, E. 2011. Information seeking behaviors of parents whose children have life-threatening illnesses. *Pediatric blood & cancer* 56, 5, 805–811.
- KNAPP, C., MADDEN, V., WANG, H., SLOYER, P., AND SHENKMAN, E. 2011. Internet use and ehealth literacy of low-income parents whose children have special health care needs. *Journal of medical Internet research* 13, 3, e75.
- KOCH-WESER, S., BRADSHAW, Y. S., GUALTIERI, L., AND GALLAGHER, S. S. 2010. The internet as a health information source: findings from the 2007 health information national trends survey and implications for health communication. *Journal of health communication* 15, sup3, 279–293.
- KUMMERVOLD, P., CHRONAKI, C., LAUSEN, B., PROKOSCH, H.-U., RASMUSSEN, J., SANTANA, S., STANISZEWSKI, A., AND WANGBERG, S. 2008. ehealth trends in europe 2005-2007: a population-based survey. *Journal of medical Internet research* 10, 4, e42.
- LEE, U., LIU, Z., AND CHO, J. 2005. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*. ACM, 391–400.
- LI, X. 2010. Understanding the semantic structure of noun phrase queries. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1337–1345.
- LI, X., WANG, Y.-Y., AND ACERO, A. 2008. Learning query intent from regularized click graphs. In

- Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 339–346.
- LI, Y., ZHENG, Z., AND DAI, H. K. 2005. Kdd cup-2005 report: Facing a great challenge. *ACM SIGKDD Explorations Newsletter* 7, 2, 91–99.
- LIU, Y., ZHANG, M., RU, L., AND MA, S. 2006. Automatic query type identification based on click through information. In *Information Retrieval Technology*. Springer, 593–600.
- LORENCE, D. P., PARK, H., AND FOX, S. 2006. Assessing health consumerism on the web: a demographic profile of information-seeking behaviors. *Journal of medical systems* 30, 4, 251–258.
- LU, C.-J. 2012. Accidental discovery of information on the user-defined social web: A mixed-method study. Ph.D. thesis, University of Pittsburgh.
- LUO, G., TANG, C., YANG, H., AND WEI, X. 2008. Medsearch: a specialized search engine for medical information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 143–152.
- MACLEAN, D. L. AND HEER, J. 2013. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of the American Medical Informatics Association* 20, 6, 1120–1127.
- MAKVANA, K. AND SHAH, P. 2014. A novel approach to personalize web search through user profiling and query reformulation. In *Data Mining and Intelligent Computing (ICDMIC), 2014 International Conference on*. IEEE, 1–10.
- MASSOUDI, K., TSAGKIAS, M., DE RIJKE, M., AND WEERKAMP, W. 2011. Incorporating query expansion and quality indicators in searching microblog posts. In *Advances in information retrieval*. Springer, 362–367.

- MATTHIJS, N. AND RADLINSKI, F. 2011. Personalizing web search using long term browsing history. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 25–34.
- MCCALLUM, A. 1999. Multi-label text classification with a mixture model trained by em. In *AAAI99 workshop on text learning*. 1–7.
- MCCAY-PEET, L. AND TOMS, E. G. 2010. The process of serendipity in knowledge work. In *Proceedings of the third symposium on Information interaction in context*. ACM, 377–382.
- MURRAY, E., LO, B., POLLACK, L., DONELAN, K., CATANIA, J., WHITE, M., ZAPERT, K., AND TURNER, R. 2003. The impact of health information on the internet on the physician-patient relationship: patient perceptions. *Archives of Internal Medicine* 163, 14, 1727–1734.
- NAAMAN, M., BOASE, J., AND LAI, C.-H. 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 189–192.
- NAGARAJAN, M., GOMADAM, K., SHETH, A. P., RANABAHU, A., MUTHARAJU, R., AND JADHAV, A. 2009. *Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences*. Springer.
- NANDA, A., OMANWAR, R., AND DESHPANDE, B. 2014. Implicitly learning a user interest profile for personalization of web search using collaborative filtering. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*. Vol. 2. IEEE, 54–62.
- NATARAJAN, K., STEIN, D., JAIN, S., AND ELHADAD, N. 2010. An analysis of clinical queries in an electronic health record search utility. *International journal of medical informatics* 79, 7, 515–522.

- NICHOLAS, D., HUNTINGTON, P., GUNTER, B., WITHEY, R., AND RUSSELL, C. 2003. The british and their use of the web for health information and advice: a survey. In *Aslib Proceedings*. Vol. 55. MCB UP Ltd, 261–276.
- NIELSEN-BOHLMAN, L., PANZER, A., HAMLIN, B., AND KINDIG, D. 2004. Institute of medicine. health literacy: a prescription to end confusion. committee on health literacy, board on neuroscience and behavioral health.
- OCAMPO, A. J., CHUNARA, R., AND BROWNSTEIN, J. S. 2013. Using search queries for malaria surveillance, thailand. *Malaria journal* 12, 1, 1–6.
- PANAHAZAR, M., TASLIMITEHRANI, V., JADHAV, A., AND PATHAK, J. 2014. Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases. In *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 790–795.
- PEREIRA, F., TISHBY, N., AND LEE, L. 1993. Distributional clustering of english words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 183–190.
- PHAN, N., BAILEY, P., AND WILKINSON, R. 2007. Understanding the relationship of information need specificity to search query length. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 709–710.
- PRATT, W. AND FAGAN, L. 2000. The usefulness of dynamically categorizing search results. *Journal of the American Medical Informatics Association* 7, 6, 605–617.
- PRATT, W. AND WASSERMAN, H. 2000. Querycat: automatic categorization of medline queries. In *Proceedings of the AMIA symposium*. American Medical Informatics Association, 655.
- PU, H.-T., CHUANG, S.-L., AND YANG, C. 2002. Subject categorization of query terms for exploring web users' search interests. *Journal of the American Society for Information Science and Technology* 53, 8, 617–630.

- QIAN, Y., SAKAI, T., YE, J., ZHENG, Q., AND LI, C. 2013. Dynamic query intent mining from a search log stream. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 1205–1208.
- RADLINSKI, F. AND DUMAIS, S. 2006. Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 691–692.
- RADLINSKI, F., SZUMMER, M., AND CRASWELL, N. 2010. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World wide web*. ACM, 1171–1172.
- ROBERTS, R. M. 1989. Serendipity: Accidental discoveries in science. *Serendipity: Accidental Discoveries in Science*, by Royston M. Roberts, pp. 288. ISBN 0-471-60203-5. Wiley-VCH, June 1989. 1.
- ROSE, D. E. AND LEVINSON, D. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*. ACM, 13–19.
- ROTO, V. 2006. Web browsing on mobile phonescharacteristics of user experience web browsing on mobile phonescharacteristics of user experience doctoral dissertation. *Technology* 152, 3, 86.
- SADASIVAM, R. S., KINNEY, R. L., LEMON, S. C., SHIMADA, S. L., ALLISON, J. J., AND HOUSTON, T. K. 2013. Internet health information seeking is a team sport: analysis of the pew internet survey. *International journal of medical informatics* 82, 3, 193–200.
- SADIKOV, E., MADHAVAN, J., WANG, L., AND HALEVY, A. 2010. Clustering query refinements by user intent. In *Proceedings of the 19th international conference on World wide web*. ACM, 841–850.
- SAVOVA, G. K., MASANZ, J. J., OGREN, P. V., ZHENG, J., SOHN, S., KIPPER-SCHULER, K. C., AND CHUTE, C. G. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): ar-

- chitecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17, 5, 507–513.
- SCHAPIRE, R. E. AND SINGER, Y. 2000. Boostexter: A boosting-based system for text categorization. *Machine learning* 39, 2, 135–168.
- SEEDOR, M., PETERSON, K. J., NELSEN, L. A., COCOS, C., MCCORMICK, J. B., CHUTE, C. G., AND PATHAK, J. 2013. Incorporating expert terminology and disease risk factors into consumer health vocabularies. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. NIH Public Access, 421.
- SHEN, D., PAN, R., SUN, J.-T., PAN, J. J., WU, K., YIN, J., AND YANG, Q. 2006. Query enrichment for web-query classification. *ACM Transactions on Information Systems (TOIS)* 24, 3, 320–352.
- SHEN, D., SUN, J.-T., YANG, Q., AND CHEN, Z. 2006. Building bridges for web query classification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 131–138.
- SHETH, A., AVANT, D., AND BERTRAM, C. 2001. System and method for creating a semantic web and its applications in browsing, searching, profiling, personalization and advertising. US Patent 6,311,194.
- SHETH, A., BERTRAM, C., AVANT, D., HAMMOND, B., KOCHUT, K., AND WARKE, Y. 2002. Managing semantic content for the web. *Internet Computing, IEEE* 6, 4, 80–87.
- SHETH, A., JADHAV, A., KAPANIPATHI, P., LU, C., PUROHIT, H., SMITH, G. A., AND WANG, W. 2014. Twitris: A system for collective social intelligence. In *Encyclopedia of Social Network Analysis and Mining*. Springer, 2240–2253.
- SHETH, A. P., PUROHIT, H., JADHAV, A. S., KAPANIPATHI, P., AND CHEN, L. 2010. Understanding events through analysis of social media.

- SIEG, A., MOBASHER, B., AND BURKE, R. 2007. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 525–534.
- SILVERSTEIN, C., MARAIS, H., HENZINGER, M., AND MORICZ, M. 1999. Analysis of a very large web search engine query log. In *ACM SIGIR Forum*. Vol. 33. ACM, 6–12.
- SMITH, C. A. 2007. Nursery, gutter, or anatomy class? obscene expression in consumer health. In *Proc AMIA Symp*. 676–80.
- SMITH, G. A., SHETH, A. P., JADHAV, A. S., PUROHIT, H., CHEN, L., COONEY, M., KAPANIPATHI, P., ANANTHARAM, P., KONERU, P., AND WANG, W. 2012. Twitris+: Social media analytics platform for effective coordination.
- SONI, S. 2015. Domain specific document retrieval framework on near real-time social health data. Ph.D. thesis, Wright State University.
- SPERETTA, M. AND GAUCH, S. 2004. Personalizing search based on user search history. submitted to cikm 04.
- SPINK, A., WOLFRAM, D., JANSEN, M. B., AND SARACEVIC, T. 2001. Searching the web: The public and their queries. *Journal of the American society for information science and technology* 52, 3, 226–234.
- SPYROMITROS, E., TSOUMAKAS, G., AND VLAHAVAS, I. 2008. An empirical study of lazy multilabel classification algorithms. In *Artificial Intelligence: Theories, Models and Applications*. Springer, 401–406.
- STEVENSON, F. A., KERR, C., MURRAY, E., AND NAZARETH, I. 2007. Information from the internet and the doctor-patient relationship: the patient perspective—a qualitative study. *BMC Family Practice* 8, 1, 1.

- STEVENSON, M. AND WILKS, Y. 2003. Word sense disambiguation. *The Oxford Handbook of Computational Linguistics*, 249–265.
- SUCHANEK, F. M., KASNECI, G., AND WEIKUM, G. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 697–706.
- SULMASY, D. P. AND SUGARMAN, J. 2001. The many methods of medical ethics (or, thirteen ways of. *Methods in medical ethics*, 3.
- TEEVAN, J., RAMAGE, D., AND MORRIS, M. R. 2011. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 35–44.
- TOUTANOVA, K., KLEIN, D., MANNING, C. D., AND SINGER, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—Volume 1*. Association for Computational Linguistics, 173–180.
- TSOUMAKAS, G. AND KATAKIS, I. 2006. Multi-label classification: An overview. *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*.
- TSOUMAKAS, G. AND VLAHAVAS, I. 2007. Random k-labelsets: An ensemble method for multilabel classification. In *Machine learning: ECML 2007*. Springer, 406–417.
- USTINOVSKIY, Y. AND SERDYUKOV, P. 2013. Personalization of web-search using short-term browsing context. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 1979–1988.
- VYDISWARAN, V. V., MEI, Q., HANAUER, D. A., AND ZHENG, K. 2014. Mining consumer health vocabulary from community-generated text. In *AMIA Annual Symposium Proceedings*. Vol. 2014. American Medical Informatics Association, 1150.

- WANGBERG, S. C., ANDREASSEN, H. K., PROKOSCH, H.-U., SANTANA, S. M. V., SØRENSEN, T., AND CHRONAKI, C. E. 2008. Relations between internet use, socio-economic status (ses), social support and subjective health. *Health promotion international* 23, 1, 70–77.
- WEAVER III, J. B., MAYS, D., WEAVER, S. S., HOPKINS, G. L., EROGLU, D., AND BERNHARDT, J. M. 2010. Health information-seeking behaviors, health indicators, and health risks. *American journal of public health* 100, 8, 1520–1525.
- WEN, J.-R., NIE, J.-Y., AND ZHANG, H.-J. 2001. Clustering user queries of a search engine. In *Proceedings of the 10th international conference on World Wide Web*. acm, 162–168.
- WHITE, R. AND HORVITZ, E. 2013. From web search to healthcare utilization: privacy-sensitive studies from mobile data. *Journal of the American Medical Informatics Association* 20, 1, 61–68.
- WHITE, R. W., BENNETT, P. N., AND DUMAIS, S. T. 2010. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 1009–1018.
- WHITE, R. W. AND DRUCKER, S. M. 2007. Investigating behavioral variability in web search. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 21–30.
- WHITE, R. W., DUMAIS, S. T., AND TEEVAN, J. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM, 132–141.
- WHITE, R. W. AND HORVITZ, E. 2009a. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems (TOIS)* 27, 4, 23.
- WHITE, R. W. AND HORVITZ, E. 2009b. Experiences with web search on medical concerns and self diagnosis. In *AMIA*.

- WHITE, R. W. AND HORVITZ, E. 2014. From health search to healthcare: explorations of intention and utilization via query logs and user surveys. *Journal of the American Medical Informatics Association* 21, 1, 49–55.
- WIECZORKOWSKA, A., SYNAK, P., AND RAŚ, Z. W. 2006. Multi-label classification of emotions in music. In *Intelligent Information Processing and Web Mining*. Springer, 307–315.
- YAJUAN, D., ZHIMIN, C., FURU, W., MING, Z., AND SHUM, H.-Y. 2012. Twitter topic summarization by ranking tweets using social influence and content quality. In *Proceedings of the 24th International Conference on Computational Linguistics*. 763–780.
- ZENG, Q. T. AND TSE, T. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association* 13, 1, 24–29.
- ZHANG, M.-L. AND ZHOU, Z.-H. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *Knowledge and Data Engineering, IEEE Transactions on* 18, 10, 1338–1351.
- ZHANG, M.-L. AND ZHOU, Z.-H. 2007. MI-knn: A lazy learning approach to multi-label learning. *Pattern recognition* 40, 7, 2038–2048.
- ZHANG, M.-L. AND ZHOU, Z.-H. 2014. A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on* 26, 8, 1819–1837.
- ZHANG, Y. AND FU, W.-T. 2011. Designing consumer health information systems: what do user-generated questions tell us? In *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems*. Springer, 536–545.
- ZHANG, Y., WANG, P., HEATON, A., AND WINKLER, H. 2012. Health information searching behavior in medlineplus and the impact of tasks. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, 641–650.

- ZHOU, K., CUMMINS, R., LALMAS, M., AND JOSE, J. M. 2012. Evaluating reward and risk for vertical selection. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2631–2634.
- ZIELSTORFF, R. D. 2003. Controlled vocabularies for consumer health. *Journal of biomedical informatics* 36, 4, 326–333.