QUALITY ORIENTED INFORMATION RETRIEVAL AND TIMELINESS ANALYSIS ON DIABETES WEBSITES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS

OF

THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

RAHİME BELEN SAĞLAM

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY

IN

THE DEPARTMENT OF INFORMATION SYSTEMS

JUNE 2014

QUALITY ORIENTED INFORMATION RETRIEVAL AND
TIMELINESS ANALYSIS ON DIABETES WEBSITES

Submitted by **Rahime BELEN SAĞLAM** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Information Systems**, **Middle East Technical University** by,

Prof. Dr. Nazife Baykal ________________________
Director, **Informatics Institute**

Prof. Dr. Yasemin Yardımcı Çetin ________________________
Head of Department, **Information Systems**

Assist. Prof. Dr. Tuğba Taşkaya Temizel ________________________
Supervisor, **Information Systems**

Examining Committee Members:

Prof. Dr. İnci Batmaz ________________________
Statistics, METU

Assist. Prof. Dr. Tuğba Taşkaya Temizel ________________________
Information Systems, METU

Assoc. Prof. Dr. Aysu Betin Can ________________________
Information Systems, METU

Assoc. Prof. Dr. Hasan Oğul ________________________
Computer Engineering, Başkent University

Assoc. Prof. Dr. Sevgi Özkan Yıldırım ________________________
Information Systems, METU

Date: **30/06/2014**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name	**:**	**RAHİME BELEN SAĞLAM**
Signature	**:**	

ABSTRACT

QUALITY ORIENTED INFORMATION RETRIEVAL AND TIMELINESS ANALYSIS ON DIABETES WEBSITES

BELEN SAĞLAM, Rahime
Ph. D., Department of Information Systems
Supervisor: Assist. Prof. Dr. Tuğba Taşkaya Temizel

June 2014, 74 pages

The foremost requirement of health information seekers is to retrieve high quality and up-to-date information from web search engine results. Current techniques rely heavily on Web graph structure and they are domain independent solutions. However, in health domain, to ensure information quality, a search engine should return results that are not only relevant to submitted query but also in accordance with evidence based medical guidelines. The aim of this thesis is to propose an automated framework which is able to retrieve high quality and up-to date health information according to evidence based medicine. The contributions of the thesis are twofold. The first one is a method which is developed to differentiate high quality and unbiased health information content from low quality ones and makes use relevance feedback, information retrieval and opinion mining techniques. The second one is a method which is developed to automatically assess the timeliness of a health web site using evidence based clinical practice guidelines. The experiments are conducted on diabetes web sites and the results show that the first method achieves 76% accuracy in detecting high quality web sites and the second method accomplishes 77% in detecting the timeliness of web sites.

Keywords: Information Retrieval, Relevance Feedback, Diabetes, Information Quality, Timeliness Analysis

ÖZ

DİYABET WEBSİTELERİNDE KALİTE ODAKLI BİLGİ ERİŞİMİ VE BİLGİ GÜNCELLİĞİ ANALİZİ

BELEN SAĞLAM, Rahime
Doktora, Bilişim Sistemleri Bölümü
Tez Yöneticisi: Yrd. Doç. Dr. Tuğba Taşkaya Temizel

Haziran 2014, 74 sayfa

Sağlık bilgisi araştıranların en önde gelen ihtiyacı yüksek kalitede ve güncel bilgiyi arama motoru sonuçlarından elde etmektir. Mevcut teknikler ağırlıklı olarak Web çizge yapısına dayanır ve ilgi alanından bağımsız çözümlerdir. Ancak sağlık alanında, bilgi kalitesini sağlamak için, bir arama motoru girilen sorguyla sadece ilgili olan değil kanıta dayalı tıp ilkeleriyle uyumlu olan sonuçları dönmelidir. Bu tezin amacı kanıta dayalı tıbba göre yüksek kalitede ve güncel bilgileri dönebilen otomatikleştirilmiş bir yapı önermektir. Tezin iki temel katkısı vardır. İlki yüksek kalite ve yansız sağlık bilgisi içeriğini düşük kalite olanlardan ayırmak için geliştirilen yöntemdir ve ilgililik geri bildirimi, bilgi çıkarma ve fikir madenciliği tekniklerini kullanır. İkincisi, kanıta dayalı klinik uygulama ilkelerini kullanarak sağlık websitelerinin güncelliğini otomatik olarak değerlendirmek için geliştirilen yöntemdir. Deneyler diyabet websiteleri üzerinde uygulanmıştır ve sonuçlar ilk yöntemin yüksek kalite websitelerini bulmada %76 ve ikinci yöntemin websitelerinin güncelliğini saptamakta %77 doğruluk elde ettiğini göstermektedir.

Anahtar Kelimeler: Bilgi Çıkarma, İlgililik Geribildirimi, Diyabet, Bilgi Kalitesi, Güncellik Analizi

dedicated to my beloved husband Ali Sağlam

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my supervisor, Assist Prof. Dr. Tuğba Taşkaya Temizel for her support, valuable guidance, supervision and friendly encouragement. I appreciate all her contributions of time and ideas to make my Ph.D. experience productive and stimulating.

I am grateful to Prof. Dr. İnci Batmaz, Assoc. Prof. Dr. Sevgi Özkan Yıldırım, Prof. Dr. Aysu Betin Can, and Assoc. Prof. Dr. Hasan Oğul for their valuable advices.

Thanks go to The Scientific and Technical Research Council of Turkey (TÜBİTAK) for providing scholarship during my study.

I would like to give my special thanks to my parents Süleyman Belen and Sevinç Belen; my sisters Esra Belen Demirtaş and Seda Belen for their love and support. Thanks to my beloved nephew Mehmet Poyraz Demirtaş for his smiles which enabled me to complete this work.

Thanks go to my friends Berna Bakır Batu, Pelin Bayraktar and Özden Özcan Top for supporting me throughout this thesis.

Finally, I would like to thank my husband Ali Sağlam for his long-standing support. Last but not the least; my inexpressible debt is to my unborn daughter for her understanding.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADA	:	American Diabetes Association
ATR	:	Automatic Term Recognition
CKD	:	Chronic Kidney Disease
CL	:	Candidate Lists
COM	:	Citation Opinion Mining
DARE	:	Database Of Abstracts Of Reviews Of Effectiveness
DB	:	Diabetes
EBM	:	Evidence Based Medicine
GN	:	Generic
HON	:	Health On The Net
idf	:	Inverse Document Frequency
LIWC	:	Linguistic Inquiry And Word Count
LR	:	Likelihood Ratio
NDIC	:	National Diabetes Information Clearinghouse
OANC	:	The Open American National Corpus
OMNI	:	Organizing Medical Networked Information
POS	:	Part of A Speech
RF	:	Relevance Feedback
RG	:	Reference Guidelines
RSJ	:	Robertson-Sparck Jones
tf	:	Term Frequency
TSV		Term Selection Value
URL	:	Uniform Resource Locator
VSM	:	Vector Space Model

CHAPTER 1

INTRODUCTION

Since the emergence of the Internet, online content has become one of the main sources of information relating to health issues. However due to the rapidly increasing number of health websites, information consumers need to access unbiased, accurate, relevant and up-to-date information. To obtain information, they use search engines which usually return highly-popular websites that were not created by medical experts and other professional organizations but rather by companies or patients that are biased towards a specific medication or alternative treatments. Thus, medical information seekers are often unable to judge whether the information is valid and/or of high quality. This situation sometimes results in users ceasing their treatment program prescribed by their practitioners or using medications without their practitioners' knowledge. The recent statistics demonstrate the seriousness of the problem. The Porter Novelli EuroPNStyles survey showed that 65 percent of people in Europe use internet when they want information about a medical query (Lambert, 2010). Another study revealed that in order to improve health outcomes particularly for chronic illnesses many physicians encourage patients to be more involved in their medical care by researching their condition on the Internet. However, other medical practitioners warn their patients against the misleading information on the internet which can heighten the patient anxiety and develop cyberchondria (Moyer, 2012). On the other hand, despite the fact that health information provided on the Web cannot substitute for doctors' advice, it is reported that 68% of health information seekers make decisions about health care and treatments based on the information on the Web (Fox & Rainie, 2002). Additionally, in the same study, 18% of health information seekers reported that they make decisions solely based on health information on the Web without consulting a doctor (Fox & Rainie, 2002). Therefore, the quality of health information becomes apparently life critical.

The majority of the studies in the literature have proposed manual information quality assessment guidelines. In these guidelines, several evaluation criteria such as the authors' credibility, citations, and last update time have been gathered in a form of questionnaire and then domain experts assessed websites based on these criteria. However, these methods require domain knowledge and the manual assessment of a particular web site takes a significant amount of time. They do not take into consideration the relevance of the information to a search query, and they are not evidence-based assessment techniques (Fallis & Frick, 2002)(Thorpe, Kiebzak, Chavez, Lewiecki, & Rudolph, 2006) (Sillence, 2006). Although both information relevance and quality are handled in few studies, these proposed methods are vulnerable to biased or even promotional contents which are created with commercial intent or amateurishly written (Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003). On the other hand, there is no indicator regarding the timeliness of the content except for the last update time.

Information timeliness refers to information which is sufficiently up-to-date at the time of publication and it is studied under the data freshness and also data accuracy quality dimensions (Shin, 2003)(Bouzeghoub, A framework for analysis of data freshness, 2004)(Peralta, 2006). Studies on health domain have shown that websites provide imperfect information and give recommendations that are not up to date with the recent literature although their last modified dates are quite recent.

Information seekers are generally suggested to check the last update dates and the presence of any broken links in order to gain insight about the currency of a web page (MedlinePlus Guide to Healthy Web Surfing, 2010). Although many web sites have accurate information, current update dates and no broken links, they may provide outdated information (Post & Mainous, The Accuracy of Nutrition Information on the Internet for Type 2 Diabetes, 2010). The timeliness assessment of content is a fundamental problem and it is hard for an information seeker who is often not a medical expert to determine and comment on the quality of web sites.

1.1 Research Questions

This thesis describes an automated framework designed to rank websites that are related to the diagnosis, treatment and control of diabetes. Further analyses are conducted on high quality web sites regarding the timeliness of the information. To achieve these goals three main research questions are answered.

R1: What are the indicators for biased information on the diabetes related websites and how can they be automatically detected? What is the most effective way to detect these bias indicators?

R2: Can we develop an engine to automatically detect diabetes web pages of both high relevance and high quality using evidence-based medicine?

R3: Can we develop an engine to automatically detect *timeliness* of diabetes web pages using evidence-based medicine guidelines published each year?

In this thesis, a quantitative research strategy was followed and the results generated by the framework were compared with the manual scores based on evidence based medicine (EBM) approach. EBM is defined as "the conscientious, explicit and judicious use of the current best evidence in making decisions about individual patients' care by gathering the best available external clinical evidence from systematic research" (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996). It aims to ensure that medical decisions are evidence based integrating individual clinical expertise and the best external evidence. In order to achieve this, EBM approach brings together experts who authoritatively review hundreds of studies on topics in their specialty. At the end of these reviews, list of effective treatments are provided. When these findings are disregarded by any rating tool, the results have the potential to be produced incompatible with the latest researches.

Different gold standards have been reported for different health issues. The manual scores used for the comparison in this study were obtained by a gold standard for EBM on diabetes were provided by the American Diabetes Association (ADA).

1.2 Contributions of the Thesis

The contributions of this thesis are summarized as providing;

- a document quality ranking method for diabetes websites which takes into account the content quality and information relevance of the website.

- a bias website identification method based on a lexical resource, SentiWordNet.
- a method which estimates to what time interval the content really belongs to according to a given reference guideline automatically.

In quality assessment phase, encouraging results were obtained from the study and they are better compared with the methods presented in the literature. In contrast with the previous studies, irrelevant web pages within websites were not eliminated manually and all the pages were assessed automatically in order to calculate the overall score of the website taking into consideration the relevance to diabetes of each web page. In addition, biased information was taken into account in the quality scoring in this study whilst other methods in the literature ignored this criterion. In timeliness prediction phase, it is demonstrated that a framework based on automatic term recognition, relevance feedback and information retrieval can be a valid indicator for content timeliness assessment of diabetes websites according to the evidence based medicine recommendations. In timeliness analysis, the archives of the high quality websites were utilized and timeliness prediction has been conducted both according to years and predefined year intervals. To the best of our knowledge, this study is the first of its kind in the literature.

1.3 Structure of the Thesis

The rest of the thesis is divided into five chapters. Chapter 2 reviews the literature surveying the quality criteria for health domain and presenting a review of basic information retrieval and extraction techniques. Chapter 3 focuses on applications of introduced techniques. In particular, the chapter discusses the applications of information retrieval, temporal information retrieval and opinion mining techniques on health domain and introduces a classification based quality assessment technique. The chapter also reports the limitations of the given techniques. Chapter 4 presents a method to automatically evaluate the quality of diabetes websites including biased content detection. Chapter 5 presents a method for timeliness analysis of diabetes websites. The scope and limitations of the developed techniques are discussed at Chapter 6. The generalizability of the techniques to other health and non-health topics are discussed in this chapter. Chapter 7 presents overall conclusions of the thesis and discusses future directions.

CHAPTER 2

LITERATURE REVIEW

To provide a foundation and context for the proposed framework, this chapter outlines the quality criteria including the dimensions focused in this study and provides a review of the techniques used in the proposed framework. The chapter also presents a close look into the methods used to evaluate health websites for information quality in the literature.

2.1 Quality Criteria for Health Domain

Many researchers have studied the dimensions of information quality and their indicators on Web sites regardless of the topics of the contents. In health domain, several studies have attempted to identify quality related features used in rating tools. Kim et al. conducted a survey in order to evaluate health websites by assessing 29 rating tools, 24 websites and several articles (Kim, Eng, Deering, & Maxfield, 1999). Similar studies were conducted results of which are summarized in Table 1.

Table 1: Proposed Quality Criteria for Health Domain

Year	**Author(s)**	**Detected features related with quality**
1999	P. Kim, T. R. Eng, M. Deering and A. Maxfield	• Content • Design • Aesthetics of sites • Disclosure of Authors • Sponsors • Currency Of Information • Ease of Use
2002	G. Eysenbach and C. Kohler	• Accuracy • Completeness • Readibility • Design • Disclosure • References Provided
2002	Heinke Kunst, Diederik Groot	• Source • Currency • Evidence hierarchy
2002	Don Fallis, Martin Frické	• Displaying the HONcode logo • Having an organization domain • Displaying a copyright
2004	Martin Frické, Don Fallis, PhD, Marci Jones, MD, Gianna M. Luszko, MD	• Inlinks to the main page of a site • Unbiased presentation of information
2004	Elizabeth Sillence	• Inappropriate name for the

	Pam Briggs Lesley Fishwick Peter Harris	website • Complex, busy layout • Lack of navigation aids • Pop up adverts • Too much text • Poor search facilities/indexes • Irrelevant or inappropriate content
2005	Kathleen M Griffiths, Helen Christensen	• DISCERN score
2005	E. M. Lewiecki, L. A. Rudolph, G. M. Kiebzak, J. R. Chavez and B. M. Thorpe	• (URL) suffix (.com, .edu)
2005	Claire R. McInerney and Nora J. Bird	• Site update--access interval
2006	Kendra L. Schwartz, MD, MSPH, Thomas Roe, MD, Justin Northrup, MPT, James Meza, MD, Raouf Seifeldin, MD andAnne Victoria Neale, PhD, MPH	• Endorsement of the site by a government agency or a professional organization
2008	Yasser Khazaal M.D. ,Sebastien Fernandez M.A. Sophie Cochand B.A. ,Isabel Reboh B.A. Daniele Zullino M.D	• DISCERN • HON label
2009	Caryl Barnes, Robin Harvey, Alex Wilde, Dusan Hadzi-Pavlovic, Kay Wilhelm, Philip B. Mitchell	• BWQC score • DISCERN score • Having editorial board • Affiliation to a professional organization

Some of these criteria are highly correlated with each other and they even overlap. As a remedy, researchers utilized the study of Stvilia et al. who have grouped the criteria that fall under a common topic using factor analysis (Stvilia, Mon, & Yi, 2009). Table 2 shows the information quality constructs and the criteria that underlie them reported by Stvilia et al.

Table 2: Information Quality Constructs

Construct	**Criteria**
Accuracy	Accuracy, Credibility, Reliability
Completeness	Completeness, Clarity
Authority	Authority
Usefulness	Ease of Use, Objectivity, Utility
Accessibility	Accessibility, Cohesiveness, Consistency, Volatility

In the following sections common quality criteria are discussed in further detail.

Until 2006, studies kept being conducted about the quality criteria and manual assessment of the quality. In 2006, Wang proposed the first study about automatic detecting indicators for quality of health information on the Web (Wang & Liu, 2007). The researchers proposed a tool called Automatic Indicator Detection Tool which detected the indicators of information quality dimensions. As the first step of their study, researchers chose 18 initial technical criteria about authority (Author’s name, Author’s credentials, and Author’s affiliations), source (Reference, Copyright), currency (Date of creation, Date of last update), content (Editorial review policy, Advertising policy, Disclaimer), disclosure (Statement of purpose, Privacy policy, Sponsorship), interactivity (Search, Contract us, Sitemap), and commercialization (Payment information). For each criterion, the measurable ones were tried to be automatically detected by the tool by searching the text and hyperlinks. They have determined the possible indicators and the set of values and the locations of these indicators on a website. For each criterion, the tool detects the measurable indicators automatically by parsing the text and hyperlinks.

Of course, there were indicators which could not be detected by the tool. Among the indicators, the ones whose detection accuracy and occurrence frequency were high were selected and the final model was developed based on these indicators.

It is important to note that this tool is a detection tool whose aim is to detect the existence of the quality indicators. Researchers did not make any further implementation to produce a quality score from these detected indicators. It is also worthwhile to indicate that these detected indicators give limited idea about the 'information quality' but nothing about 'information relevance'. Consequently while ranking the web pages, considering these indicators alone is not sufficient.

2.1.1 Accuracy

This criterion requires the content of information to be examined for the evidence available to support the claims (Eysenbach & Kohler, 2002). It is cited as the most important criteria in the majority of the studies (Stvilia, Mon, & Yi, 2009) (Fallis & Frick, 2002) (Marshall & Williams, 2006). Frick and Fallis (Fallis & Frick, 2002) studied the relationships between the quality indicators (e.g., a copyright sign, citations, a lack of advertising) and web page accuracy. The researchers indicate that there are three indicators which are correlated with accuracy: displaying the HONcode (Health on the Net code) logo, having an organization domain, and displaying a copyright. Despite the common assumptions, many proposed indicators such as the author being identified and the author having medical credentials are not detected as correlated with accuracy. Also lack of currency and advertising are not detected as correlated with inaccuracy. Alternative terms for the accuracy can be given as "reliability" or "conventionality of information".

2.1.2 Comprehensiveness

Completeness or comprehensiveness is another information quality criteria used by evaluators of health information on the Internet (Eysenbach & Kohler, 2002). Some empirical work shows that the more information provided, the more a Web site is trusted by consumers (Dutta-Bergman, 2004). However, it does not mean that a website is supposed to cover a broad scope of topics in order to be complete or cohesive (Eysenbach & Kohler, 2002). In a hyperlinked world, any consumer can find the complementary information mouse-click away. Consequently, from information consumer's perspective, a website is comprehensive even if it does not provide the 'complete' information but provides just links for complementary information.

2.1.3 Design

This feature which is also called as "design and aesthetics" (Kim, Eng, Deering, & Maxfield, 1999) refers to the visual aspects of a website. It also covers interactivity, appeal, presentation, use of media and graphics.

2.1.4 Ease of Use

Ease of use is an important consideration for information consumers. Even websites provide valuable sources of information, if they are not loaded fast enough or navigation is not easy, then their contribution and usefulness will be lost. While assessing the usefulness, organization of the website, ease of navigation, existence of Help or Search Tips pages and search capability are taken into account.

2.1.5 Disclosure

Disclosure which is also called authority requires a website to include information about authors, sponsors, and site owners. It also includes identification of purpose, sources of support, nature of organizations, authorship and origin (Moreno, Del Castillo, Porcel, & Herrera-Viedma, 2010). The criteria can also be expanded regarding the authority as follows; Website origin place, owner's contact address, Website sponsors and investors, and Interest conflict declaration must also be displayed.

2.1.6 References

In some studies this criterion is accepted as authority of sources which means reputation or quality of source, credibility, and trustworthiness. On the other hand, some authors use this notation to imply whether the information provided on websites includes citations of the sources.

2.1.7 Timeliness

Timeliness means currency of information which refers to date of creation, posting and amendment to the contents. This criterion is important in health domain since health domain subjects to frequent changes and requires content to be aligned with the recent research. Timeliness of the information given in the referenced websites must also be considered by the website owners while assessing this criterion.

Since online content is used as a main source of information on health issues, the timeliness of websites have become more of an issue for information seekers. The studies have shown that health recommendations on web sites vary in quality and many users fail to access the reliable and accurate information on world wide web (Post & Mainous, The Accuracy of Nutrition Information on the Internet for Type 2 Diabetes, 2010) (Scullard, Peacock, & Davies, 2010)(Morr, Shanti, Carrer, Kubeck, & Gerling, 2010)(van der Marel, et al., 2009)(Clark, 2002). Information seekers are generally suggested to check the last update dates and the presence of any broken links in order to gain insight about the currency of a web page (MedlinePlus Guide to Healthy Web Surfing, 2010). Although many web sites have accurate information, current update dates and no broken links, they may provide outdated information (Post & Mainous, The Accuracy of Nutrition Information on the Internet for Type 2 Diabetes, 2010). The timeliness assessment of content is a fundamental problem and it is hard for an information seeker who is often not a medical expert to determine and comment on the quality of web sites.

There are limited studies which have worked on different timeliness aspects of health web sites. Post and Mainous manually assessed the accuracy of nutrition information on the Internet for Type 2 diabetes and concluded that the date of when the websites were updated is not correlated with the accuracy of the provided information (Post & Mainous, The Accuracy of Nutrition Information on the Internet for Type 2 Diabetes, 2010). On the other hand, portals such as HONSearch assess the currency based on the last update time. Such a finding shows that current approaches fail to assess currency of a given content in means of evidence-based medicine. Consequently, an automated system which can support the information seekers in assessing the timeliness of a given content is important.

2.1.8 Evidence-Based Medicine

When the quality of health websites is considered, “evidence based medicine” is an important phenomenon to understand. Oxford Center for Evidence based medicine defines EBM as ”the conscientious, explicit and judicious use of current best evidence in making decisions about care of individual patients”. It is accepted as the “gold standard” for medical treatment and can serve as a guide for several professionals and organisations such as doctors, patients or health insurance companies. In whole health domain, treatments recommended online should be consistent with EBM. In order to eachieve this, evidence-based health information is routinely disseminated to health professionals and it is aimed to assist clinical decision making. Doctors should apply treatments that are known to be effective with the scientific evidence provided. Similarly, when these guidelines are provided to consumers, they have potential to improve health outcomes guiding consumers to select effective self-help techniques (Hibbard, 2003).

For online search systems, EBM provides a very specific definition of quality. That is, a health portal should return results that are not only relevant to the query but also in accordance with evidence-based medical guidelines.

There is no generic gold standard exists which is applicable to a great number of health related web sites and evaluates the accordance with EBM. Different gold standards are reported for different health issues. One of the standards provided for diabetes is ADA Clinical Practice Recommendations (American Diabetes Association, 2011). ADA Clinical Practice Recommendations are published each year and based on a complete review of the relevant literature about diabetes. They are grouped under several sections related with diabetes. Revisions are made each year and a brief summary is given for each revision in which new sections or sections that are revisited are stated.

2.1.9 Quality Labels

In health domain, the major self-regularity agencies started to develop quality and ethical standards for health information on the Internet. In 1995, the Health on the Net (HON) foundation has developed principles named as Net Code of Conduct (HONcode) (Health on the Net Foundation, 2013). The HON is a non-profit organization which works under the auspices of the Geneva Ministry of Health in Geneva, Switzerland. Principles proposed by the foundation are authoritativeness, statement of the purpose, confidentiality, reference section, justification of claims, website content details, disclosure of funding resources, and advertising policy. The websites which display the logo of the foundation indicate that they follow the HONcode principles. The HONCode is the oldest and the best known quality label on the Web adopted by several websites. While rating the quality, it uses eight criteria; indication of the authorship, purpose of the website, confidentiality of user’s information, references, justification of claims, contact information, disclosure, advertising policy and currency of information. To assess currency, it considers creation and modification date. HONCode checks adherence to the code by websites displaying its labels regularly.

Several companies provided similar third party rating services like HONcode logo some of which are Trustee, Hi-Ethics and the Internet Content Rating Association (ICRA) (Resnick & Miller, 1996). These studies have led the information quality assessment and still can be used as an indicator of accuracy. However, these logos are not present in most of the health related websites (Wang & Liu, 2007). Consequently, it is not appropriate to handle these logos as the only indicator of trustworthiness while evaluating the websites.

2.1.10 User Guidance System

User guidance systems consist of checklists to evaluate a given website and enable users to check if a website and its contents follow certain standards (Wang & Liu, 2007). They can be specific to a health condition or general purpose. They help information seekers to personally evaluate the health websites. Organizing Medical Networked Information (OMNI) (Norman, 1998) can be given as an example. OMNI provides detailed guidelines to be used in evaluating the websites and experts teams assess the medical information depending on these guidelines. DISCERN was given as another guideline which was developed in 2000 to facilitate information quality assessment for information seekers and information providers about treatment choices (University of Oxford, Division of Public Health and Primary Health Care, 2010). The tool provides 16 questions under three sections; the reliability of the publication, the quality of information and overall rating of the publication. However, Bernstam et al. (Bernstam, Shelton, Walji, & Meric-Bernstam, 2005) revealed that a small part of these guidelines was likely to be practically utilized by end users. Consequently, researchers tended to study methods that the end users can assess the web page on their own in a practical way. Several instruments have been proposed in which a trust logo or self-evaluation questionnaires were utilized. Among those instruments Bernstam studied the effectiveness and usability. Among the instruments, 8.7% of them have less than 10 elements that could be assessed by patients. Criteria of only 7 instruments such as date of creation, disclosure of physician’s credentials, references provided disclosure of ownership or statement of purpose can be evaluated objectively.

2.1.11 Portal (Filter)

Health portals are other online medical tools which provide evaluated high quality health information. HONSearch can be given as an example which provides links to the websites that adhere to HONCode of conduct. Advantage of such portals is that information seekers do not have to consume time and effort to judge the quality of the websites, or check the existence of a quality label. However, coverage of these portals is limited and they return webpages belonging to a small number of websites.

2.2 Information Retrieval and Extraction Techniques

This section presents an introduction to the techniques used in this thesis. It provides a review of basic information retrieval techniques, including text-based ranking techniques and relevance feedback techniques for query expansion; an overview of automatic term recognition techniques; an introduction to temporal information retrieval techniques and a detailed review of opinion mining techniques.

2.2.1 Text Based Document Ranking

As the amount of textual information available in electronic form increases, efficient and effective text retrieval techniques have become critical to manage them. Text-based document ranking is a principal way of managing high volume documents providing a search system. It typically assigns scores to documents based on the distribution of query terms within both the document and the corpus. A query term can be an individual word, (n-)tuples of words or a word with its synonyms. There are three main types of ranking models: Boolean matching, vector space model (VSM) and probabilistic model. Models, except for the Boolean model normalize term contribution using document length. Boolean matching clearly identifies which documents should be retrieved. However since it does not provide any ranking by degree of match, it has become less popular in Web search.

Since the effectiveness of VSM and probabilistic models are reported to be similar, there is no reason for selecting one over the other.

Probabilistic ranking function Okapi BM25 has been applied in the experiments within this thesis.

Boolean Model

In Boolean exact match retrieval models, a document is estimated to be either relevant or non-relevant to a query disregarding the degree of match. The function is;

$$S\ \ D,Q\ \ =\ \ \begin{matrix} 1\ \ if\ Q \subset D \\ 0\ \ otherwise \end{matrix} \tag{1}$$

where *D* denotes a document and *Q* is a query.

Vector Space Model

In this model, a document is represented by a vector in a multi-dimensional space with an orthogonal dimension for each term in the corpus (Salton, Wong, & Yang, 1975). A distance function which calculates the degree of overlap between the query and the document is used to estimate the degree of relevance of a document to a query. The vector can be made up of 1's and 0's indicating the presence or absence of a term in the document. Term and document statistics can be used as well to construct the vector. Two important terms regarding document statistics are term frequency (tf) and inverse document frequency (idf) (Jones, 1972). Term frequency is the number of times a term appears in the documents. It is assumed that a document that contains a term more often is more likely to be relevant to a query having that term. According to Sparck-Jones, the potential for a term to discriminate between candidate documents is an important statistics and can be computed by dividing the total number of documents by the number of documents containing the term which measures whether the term is common or rare across all documents.

Using two measures, an optimal weight is given to a term by combining the definitions of term frequency and inverse document frequency, to produce a composite weight for each term in each document as follows;

$$w_{t,D} = tf_{t,D} * idf_t \tag{2}$$

where *idf* is:

$$idf_t = \frac{N}{n_t} \tag{3}$$

where *N* is the number of documents in the corpus, and n_t is the number of documents in the corpus that contain term *t*. Thus the idf of a rare term is high, whereas the idf of a frequent term is likely to be low.

The $w_{t,D}$ will be high when t occurs many times within a small number of documents whereas it will be lower when the term occurs fewer times in a document, or occurs in many documents; and finally lowest when the term occurs in virtually all the documents.

There are several functions to compute the distance between a document vector and a query vector. One of the most common functions is cosine measure of similarity which is calculated as;

$$S\ \ D,Q\ \ = \frac{D * Q}{D\ * \ Q} \tag{4}$$

where D and Q represent document and query vectors respectively and D and Q represent the length of the vectors.

Probabilistic Models

In probabilistic retrieval models, the probability of a document being relevant to a query is estimated (Robertson & Jones, Relevance weighting of search terms, 1976) and similar to vector-space models, $tf_{t,D}idf_t$is utilized while computing the probability. Okapi BM25 also considers document length for normalization since *tf* values tend to be larger in longer documents and it causes a potential bias that longer documents will be assigned to higher scores. In order to compensate that effect, Robertson derived the document length normalization. Document length normalization is important especially for the corpora in which the length of documents highly differ. Web is a good example of these corpora.

Okapi BM25 computes the score for each word or word phrase in the query as follows;

$$\mathrm{w}_{\mathrm{t,D}}^{\mathrm{BM25}} = Q_{wt} * tf_{t,D} * \frac{\log(\frac{N-n_t+0.5}{n_t+0.5})}{2 *\ \ 0.25 + 0.75 * \frac{ld}{avdl}\ \ +\ tf_{t,D}} \tag{5}$$

where Q_{wt} is the weight of the term *t* by the query. *N* is the total number of documents. In the corpus, n_t is the number of documents containing *t*, and $\log(\frac{\mathrm{N}-\mathrm{n_t}+0.5}{\mathrm{n_t}+0.5})$ is the idf value of the term *t*. *ld* implies the length of the documents and *avdl* stands for the average length of documents in the corpus.

The final score for a document is the sum of all the term weights:

$$S^{BM25}(D,Q) = \underset{t\epsilon Q}{\ \ } w_{t,D} \tag{6}$$

Relevance Feedback

Relevance Feedback (RF) is an information retrieval approach of “query by example”. The aim is to find relevant documents given the example sets of relevant and irrelevant documents (Rocchio, 1971).

In the relevance feedback approach, term frequency distributions in relevant and irrelevant documents are compared and queries are generated which consist of words and phrases with appropriate weights. Here, the assumption is that terms in relevance query occur frequently in the relevant text but rarely in the irrelevant ones. The generated queries are the input of the text retrieval systems that compute relevance scores for documents.

RF techniques utilize several models of retrieval containing vector space models and probabilistic models discussed in section 2.2.1. In this thesis, a probabilistic model introduced by Robertson and Sparck Jones (RSJ) has been applied in which probability of terms in relevant and non-relevant documents are taken into account (Robertson & Jones, Relevance weighting of search terms, 1976).

Important definitions utilized in RSJ formula are summarized in Table 3:

Table 3: Definitions for RSJ formula

Equation		**Definition**
n	=	Number of documents that contain the term
r	=	Number of relevant documents that contain the term
n-r	=	Number of non-relevant documents that contain the term
R	=	Number of relevant documents
N	=	Number of documents in the collection
R-r	=	Number of relevant documents that do not contain the term
N-n	=	Number of documents that do not contain the term
N-R	=	Number of non-relevant documents
N-n-R+r	=	Number of non-relevant documents that do not contain the term

Given the definitions, the probabilities are defined as follows;

Table 4: Components of RSJ formula

Probability Equation		**Definition**
r/R	=	probability that a relevant document contains the term
(n – r)/(N-R)	=	probability that a non-relevant document contains the term
n/N	=	probability that a document contains the term
(R – r)/R	=	probability that a relevant document does not contain the term
(N – n – R + r)/(N – R)	=	probability that a non-relevant document does not contain the term
(N – n)/N	=	probability that a document does not contain the term
r/(R – r)	=	odds that a relevant document contains the term
(n – r)/(N – n – R + r)	=	odds that a non-relevant document contains the term
n/(N – n)	=	odds that a document contains the term

Finally, Robertson-Sparck Jones formula is as follows;

$$F = \log \frac{((r+0.5)/(R-r+0.5))}{((n-r+0.5)/(N-n-R+r+0.5))} \tag{7}$$

2.2.2 Automatic Term Recognition Techniques

In this section, automatic term recognition techniques (ATR) which are utilized by the proposed framework are introduced. In the literature, ATR techniques are required for several reasons such as thesaurus construction, knowledge organization or automatic keyword extraction. The methods used for ATR can be classified according to its use of background knowledge, i.e. a corpus in a general domain. Termhood and unithood are two characteristics that can be used for classification. Termhood is measured as the degree by which a linguistic unit (word) is related to the domain-specific concept and computed based on the frequency of occurrence (Kageura & Umino, 1996). Weirdness is a termhood-based method which is based on the assumption that distribution of terms in a specialized corpus (domain) and in a general corpus (background) differ significantly from each other(Ahmad, Gillam, & Tostevin, 1999). It is expressed by the following formula:

$$\text{Weirdness}\ i\ = \frac{\frac{f_s(i)}{n_s}}{\frac{f_g(i)}{n_g}} \qquad (8)$$

where $f_s\ \ i$ and $f_g(i)$ are the frequencies of word *I* in the specialized and the general corpus respectively, n_s is the total number of words in specialized corpus and n_g is the total number of words in the general one. Since the original Weirdness was defined for single words, in this thesis, geometric average of Weirdness values of each word in the term is computed for multi-word terms as done in (Knoth, Schmidt, Smrz, & Zdrahal, Towards a framework for comparing automatic term recognition methods, 2009). Likelihood Ratio (LR) is similar to Weirdness. However, significance of differences between word frequencies in the domain and those in the background corpus is measured (Manning & Schütze, 1999).

$$\text{p} = \frac{f_s + f_g}{n_s + n_g} \qquad \text{p}_s = \frac{f_s}{n_s} \qquad \text{p}_g = \frac{f_g}{n_g}$$

$$LR = \log L(f_s, n_s, p) + \log L(f_g, n_g, p) - \log L(f_s, n_s, p_s) - \log L(f_g, n_g, p_g) \qquad (9)$$

where $L(k, n, x) = x^k(1-x)^{n-k}$

Unithood is used for compound terms that consist of more than one word and gives the collocation strength of the words contained (Zhang, Iria, Brewster, & Ciravegna, A comparative evaluation of term recognition algorithms, 2008). While determining the unithood, the significance of the co-occurring words is measured. C-Value is a multiword ATR method which is an example for unithood-based methods focusing on the nested terms. The motivation is that a candidate term should occur frequently on its own, not nested in the other candidate terms (Korkontzelos, Klapaftis, & Manandhar, 2008). The method covers three main principals; the terms that appear the most are extracted, the nested terms which are covered within longer terms are penalized and the number of the words candidates consist of is taken into account. The background knowledge is ignored in this method.

$$C-value\ \ a\ = \begin{array}{ll} log_2\ a\ * f(a), & if\ a\ is\ not\ nested \\ log_2\ a\ *\ \ f\ \ a\ -\frac{1}{|T_a|}\ \ _{b\in T_a}\ f(b) & ,otherwise \end{array} \qquad (10)$$

In Equation 10, a and b are the candidate terms where f denotes the frequency and T_a is the set of candidate terms which contain a.

Glossex Method (Kozakov, et al., 2004) is based on two measures; the first of which evaluates the degree of domain specificity (TD) that is equal to Weirdness and the second of which considers the idea of term cohesion of multiword terms. In term cohesion calculation, the aim is to measure the association of an arbitrary n-gram (n>1), and to give higher values to terms having high co-occurrence frequencies(Kozakov, et al., 2004). The term cohesion can be expressed by the following formula:

$$TC_{D_i}\ t\ = \frac{n * tf_{t,D_i} * logtf_{t,D_i}}{{}_{j=0}^{n} tf_{w_j,D_i}} \tag{11}$$

where n is the number of words included in the term t, w$_j$ is a jth word in term t. The two measures are combined by two adjustable coefficients. In this study these two measures were given the same weight as 0.5.

$$GlossEx = \alpha * TD\ t\ + \beta * TC(t) \tag{12}$$

Glossex Method has been reported as a superior method in term extraction in the literature(Zhang, Iria, Brewster, & Ciravegna, A comparative evaluation of term recognition algorithms, 2008).

2.2.3 Temporal Information Retrieval Techniques

Timeliness and Currency Revisited

Information timeliness refers to information which is sufficiently up-to-date at the time of publication and it is studied under data freshness and also data accuracy quality dimensions (Shin, 2003)(Bouzeghoub, A framework for analysis of data freshness, 2004)(Peralta, 2006). From a user's point of view, it has two sub-dimensions: currency and timeliness. Currency is estimated as the difference between data extraction time and the data delivery time and commonly used in data warehousing systems (Theodoratos & Bouzeghoub, 1999). Timeliness measures the extent to which the age of the data is appropriate for the corresponding task (Wang & Strong, 1996).

Depending on the objectives of the applications and where it is used, data freshness has several definitions and different metrics for measurement (Bouzeghoub, A framework for analysis of data freshness, 2004). Intuitively, it considers whether the information is fresh enough with respect to the user expectations (Peralta, 2006). On the other hand timeliness measures the extent to which the age of the data is appropriate for the corresponding task (Wang & Strong, 1996). In web systems, it is also related to data volatility which denotes the time interval in which data has remained valid (Gertz, Özsu, Saake, & Sattler, 2004). Consequently, timeliness is accepted to be highly related with accuracy (Peralta, 2006). A datum is outdated at time *t* if it is incorrect at *t* but was correct before *t*. Semantic correctness deviation metric measures the semantic distance between a system datum and its real-world correspondent datum in which calculation differs depending on the data type and the application. For instance for numeric data, it can be calculated as the difference between values (Shankaranarayan, Ziad, & Wang, 2003) or for string data, the number of characters that changed can be considered (Navarro, 2001). For more complicated cases, the comparison is carried out against a reference value which could be synthesized from other data sources.

Temporal Information Retrieval

Temporal information retrieval combines temporal relevance with document relevance and aims to return *temporally relevant documents*. In a well-known study, Alonso et al. aimed to extract temporal information from the documents and clustered them along a timeline supporting multiple time granularities using named-entity extraction (Alonso, Gertz, & Baeza-Yates, 2009)(Alonso, Strotgen, Baeza-Yates, & Gertz, 2011).

Temporal entities in a text, extracted using named-entity extraction techniques, appear as sequences of tokens or words whether explicitly or implicitly. For instance “October 10 2013” is an explicit temporal expression whereas “Columbus Day 2008” is an implicit temporal expression which needs to be mapped to an explicit expression such as “October 12 2008”. Some temporal entities may also appear as relative temporal expressions that need to be referenced to another explicit or implicit expression to be anchored in a timeline. For example, the expressions “today”, “on Thursday” or “next week” can only be mapped to a date if the document’s creation date is known and used as a reference.

A time-aware document ranking methodology relying on time-aware query suggestions is proposed by Miyanishi and Sakai (Miyanishi & Sakai, Time-aware structured query suggestion, 2013). Their study makes suggestions along a timeline and helps users access relevant web pages. There are also studies which date a document based on temporal language model (Kanhabua & Norvag, Using temporal language models for document dating, 2009). In this approach, time partition of a document is found based on the overlapping term usage in the documents. For instance, if a document contains the word “tsunami”, the corpus statistics are checked accordingly and the time partition that the word “tsunami” was used most frequently is assigned to the document. In the study of Lin et al. (Lin, Chen, & Brown, MedTime: A temporal information extraction system for clinical narratives, 2013), it is aimed to construct patient’s clinical timeline from the text.

Blog distillation or blog feed search is another area where temporal relevance feedback is applied. Blog distillation aims to rank blogs according to their recurring central interest to the topic of a user's query. It is different from the aforementioned studies since it is related to the temporal properties of blogs and topics. The problem arises since distillation queries are often multifaceted and vocabulary usage in the relevant documents to a query may change over time expressing different aspects of the query. Consequently, a term in a query may not be a good indicator of the query topic in all different time intervals. Keikha et al. focused on temporal properties of blogs and proposed a time-based query expansion method which expands queries with different terms at different times to overcome this term mismatch problem (Keikha, Gerani, & Crestani, 2011). They generated one query for each time point and utilized the top N terms for each day using the KL-divergence between the term distribution of the day and the whole collection in the daily queries. Here, they did not consider the timeliness aspects of the information quality but focused on assessing relevance effectively with time-aware queries.

Temporal distribution of documents were also used in a study conducted by Peetz et al. (Peetz, Meij, de Rijke, & Weerkamp, 2012) in news corpora which is inherently temporal. In the study, it is aimed to exhibit bursts with the assumption that documents in the bursts are more likely to be relevant. Here, researchers define a burst to be a time period where unusually many documents are published. As the bursts are detected based on (pseudo-)relevance feedback, the queries are updated with the most distinguishing terms in high-quality documents sampled from bursts. It is again aimed to retrieve most relevant documents using temporal relevance feedback techniques.

2.2.4 Opinion Mining and Sentiment Analysis

The term *opinion mining* first appeared in a paper by Dave et al. that was published in the proceedings of the 2003 WWW conference (Dave, Lawrence, & Pennock, 2003). In this study Dave et al. developed a method for automatically distinguishing between positive and negative reviews within the product reviews on the Web. However, the term has recently been interpreted as more broadly which covers different analysis types of evaluative text. The history of the research area *sentiment analysis* is parallel to *opinion mining* in certain respects. In 2001 Das and Chen (Das & Chen, 2001) and Tong (Tong, 2001), who were interested in analyzing market sentiment from stock message boards, used the term "*sentiment*" in reference to the automatic analysis of text to be evaluated and tracking of the predictive judgments.

There are many studies that utilize natural language processing techniques or machine learning techniques to classify reviews according to their polarity (either positive or negative) (Turney, 2002) (Pang, Lee, & Vaithyanathan, 2002) (Yi, Nasukawa, Bunescu, & Niblack, 2003).

The studies regarding opinion mining can be classified as applications for review related websites, applications as sub-component technology, applications in business and government intelligence and applications across different domains (Pang & Lee, 2008). Summarizing user reviews and fixing user ratings can be given as examples to the first application type. There are cases where users have accidentally selected a low rating although their review indicates a positive evaluation. Also when user ratings are biased and need correction, automated classifiers are developed to provide such updates. Such cases can be fixed with the approach under this application. Recommendation systems which do not recommend items and receiving a lot of negative feedback, are examples of applications as sub-component technology. Sub-component technologies provide supportive facilities to the main component. Detection of "flames" (overly heated or antagonistic language) in email or other types of communication is another study under this topic. For example, Jin et al. proposed a method to detect webpages including sensitive content which is inappropriate for ads placement (Jin, Li, Mah, & Tong, 2007). Also Riloff et al. stated that information extraction can be improved by discarding information in subjective sentences (Riloff, Wiebe, & Phillips, 2005) that are detected by a subjective sentence classifier.

Additionally, there are studies that utilize opinion mining to determine whether an author is citing a piece of work as supporting evidence or as a research that he or she dismisses. Such a research area is named as citation opinion mining (COM) and is based on existing semantic lexical resources and NLP tools, aiming to create a network of opinion polarity relations between documents and citations.

Politics is another application area of opinion mining. Understanding what voters are thinking (Goldberg, Zhu, & Wright, 2007) (Hopkins & King, 2007) and what public figures support or oppose, are widely studied tasks in the area (Bansal, Cardie, & Lee, 2008) (Thomas, Pang, & Lee, 2006).

CHAPTER 3

BACKGROUND

The previous chapter provided an introduction to the techniques utilized in this thesis. In this section, the implementations of the techniques introduced in the previous section for health domain are presented.

3.1 Information Retrieval Techniques on Health Domain

In the literature, relevance feedback and text based information retrieval techniques have been applied on health domain to rank health related webpages in means of evidence based medicine. In their study (Griiffiths, Tang, Hawking, & Christensen, 2005) Griffiths et al., exploited EBM in assessing health related contents providing a framework that returned results that were not only relevant to the query but also in accordance with the EBM guidelines.

In the study, the relevance feedback technique was used to derive two lists of terms; one representing relevant documents and other representing high-quality documents on the topic of *depression*. These two lists served as queries for search engines to find relevant and high quality documents. The weights of the terms in the queries were computed using the RSJ formula as given in . Here, the purpose of weighting is to assign high values to discriminating terms in high quality web pages. During the development of the quality and relevance queries, the query terms were selected by computing Term Selection Values (TSVs) for each candidate term. The terms were ranked in descending order and the terms that had a rank above a certain threshold were selected.

Using 347 documents which were previously judged to be relevant on the topic of depression and 9,000 documents with very low probability of relevance to that topic, Griffiths et al. used the 20 highest TSVs and the two-word phrases with the 20 highest TSVs and generated the relevance queries (Griffiths, Tang, Hawking, & Christensen, Automated assessment of the quality of depression websites, 2005).

The learned queries were processed over the collection using the search engine to compute relevance and quality scores for all documents. The engine returned a score for each webpage and the score of a website was computed by calculating the average score of its web pages. Okapi BM25 was utilized for the retrieval task.

The relevance score for a site was computed using equation below;

$$S_r = \alpha * \ R_{score} + \ 1 - \alpha \ \ * Norm(\ln(|R|)) \tag{13}$$

where α represents the relative degree of importance of the relevance score, R_{score},compared to the number of retrieved documents, Its values lie in the range of zero to one. $Norm(\ln(|R|))$ is the normalized value of $\ln(|R|)$ which is obtained by dividing $\ln(|R|)$ and ln R_{max} . R_{max} is the maximum number of relevant pages per site.

Similarly, the quality score for a site was calculated as follows;

$$S_q = \alpha * \; Q_{score} + \; 1 - \alpha \; * Norm(\ln(|Q|)) \tag{14}$$

Here, Q_{score} means the mean quality score obtained by search engine and $Norm(ln(|Q|))$ is the normalized value of $ln(|Q|)$, obtained by dividing $ln(|Q|)$ by $ln(|Q_{max}|)$, where Q_{max} is the maximum number of retrieved documents per site.

The queries consisted of many words and phrases that were given weights. Documents with non-zero scores were retrieved and the mean relevance and quality scores were computed for each website. The scores were normalized such that the highest score to become as 1.0.

The overall score was computed using the following equation;

$$S = \beta * S_q + \; 1 - \beta \; * S_r \tag{15}$$

where β is the relative degree of importance between the quality score and the relevance score whose value ranges between zero and one.

In Equation 15, the values for parameters α and β were chosen so that the maximum correlation was achieved between the scores that were automatically computed and the scores assigned by human raters. In order to do this, the possible values of α in steps of 0.01 were selected between zero and one. The correlations of results of the two methods were compared and its value is fixed to the value which maximizes the correlation. The correlation was computed using Pearson correlation test. The researchers reported that the highest correlation were achieved as 0.806 when α = 0.76 and β = 0.3. These learned values were used in the test phase.

In their study, Griffiths et al. collected thirty depression information websites from Yahoo, LookSmart(looksmart, 2014) and DMOZ (DMOZ open directory project, 2014) which are the three major human-compiled search engines on the World Wide Web. The depression subdirectories of the search engines were utilized.

The researchers used the guideline concerning the mental health issue of depression provided by Centre for Evidence-based Medicine at the University of Oxford. In the validation phase, 30 websites were scored by human raters using the evidence-based criteria given at this guideline and the proposed method. Then the correlation was computed by Pearson correlation test. The linear correlation between these two measures was high and significant (ρ=0.851, p<0.001, n=29, df=27).

3.2 Temporal Information Retrieval in Health Domain

There are limited number of studies which have worked on different timeliness aspects of health web sites. One of the studies that focus on temporal information extraction on health domain was conducted by Lin et al. (Lin, Chen, & Brown, 2013) in which aim was to construct patient's clinical timeline from text. In health domain, temporal information systems are generally developed to facilitate healthcare management such as predicting disease risk or progression or searching similar clinical cases (Klimov, Shahar, & Taieb-

Maimon, Intelligent visualization and exploration of time-oriented data of multiple patients, 2010). Lin et al. focused on two tasks; determining clinically relevant events from clinical narratives, and temporal expression recognition and normalization which require both the recognition of clinical temporal expressions and the retrieval of temporal information from each temporal expression. For this purpose, researchers extracted morphological, syntactic, semantic, and composite features and applied machine learning methods on them. They have extracted each feature in a different way. For instance, for morphologic features, they utilized stemmed strings, word lemma and part of a speech (POS) tags whereas for semantic features, they made use of medical abbreviations in the texts.

Post and Mainous manually assessed the accuracy of nutrition information on the Internet for Type 2 diabetes and concluded that the website update dates are not correlated with the accuracy of the provided information (Post & Mainous, The Accuracy of Nutrition Information on the Internet for Type 2 Diabetes, 2010).

3.3 Assessment of Biased Contents Using Opinion Mining

In health websites, biased information can appear for many reasons consciously or unconsciously. Although there is limited study regarding detection of biased content, there are clues of biased information suggested by domain experts to keep in mind while searching health information (MedlinePlus Guide to Healthy Web Surfing, 2010). In the guidelines, users are warned to beware of bias considering several issues. One of them is the writing style. Users are suggested to be cautious if the site uses a sensational writing style (lots of exclamation points, for example). It is also underlined in the suggestions that there is a big difference between a site that says, "I developed this site after my heart attack" and one that says, "This page on heart attack was developed by health professionals at the American Heart Association. Some of the terms that are mostly used in biased websites are counted as “breakthrough”, “secret ingredient” or “miracle”.

The most related research on opinion mining for assessment of biased contents was conducted in 2008 (Denecke, 2008). Denecke analyzed medical blogs in her study where an algorithm for classifying posts according to their information content was introduced. The method allows distinguishing informative from affective posts and provides information on medical relevance. Denecke defines affective posts as the posts which describe daily activities, ideas and feelings about treatments, diseases or medications. Posts are also accepted as affective if they do not contain any medical content. Here it is important to underline that most of the contents that are defined as affective can be considered as biased. Recall that one of the clues of biased content was stated as sensational writing style by MedlinePlus (MedlinePlus Guide to Healthy Web Surfing, 2010).

Informative costs are considered to contain general or disease specific news, experiences or research results. Denecke crawled 7 web logs for his study.

The framework proposed in this study consists of two components:

- a module to determine the proportion of medical content
- a module to determine the proportion of affective content

Features generated in two modules are later used for classification. In the first module an existing information extraction system (SeReMeD) is used to extract entities on diagnoses, procedures and medications (Denecke & Bernauer, 2007).

In the second module in order to determine the proportion of affective content, SentiWordNet (Esuli, 2006) is used. Esuli designed SentiWordNet as a lexical resource in which each synset of WORDNET (version 2.0, http://wordnet.princeton.edu) is associated to three numerical scores Obj(s), Pos(s) and Neg(s), describing how Objective, Positive, and Negative the terms contained in the synset are. These triples sum up to 1. For example for the synset of the term *bad*, the triple 0, 1, 0 (positivity, negativity, objectivity) is assigned. The assumption that caused to use synsets instead of terms is that different senses of the same term may have different opinion-related properties. Synsets may have nonzero scores for all the three categories, which would indicate that the corresponding terms in the synset have each of the three opinion-related properties only to a certain degree.

SentiWordNet has been created automatically by means of a combination of linguistic and statistic classifiers. It is freely available for research purposes, and has a Web-based graphical user interface. It has been applied in different opinion related tasks such as subjectivity analysis and sentiment analysis with promising results. The approach presented by Denecke is inspired from the method proposed by Zhang et.al.(Zhang & Zhang, 2006) who used SentiWordNet for determining subjective adjectives and estimating the probability that a document contains opinion expressions.

Other than SentiWordNet, there are some resources that can be utilized for the same requirements. SentiWords, Harvard General Inquirer or Linguistic Inquiry and Word Count (LIWC) can be given as examples. SentiWords is a licensed resource covering roughly 155.000 words associated with a sentiment score included between -1 and 1 (SentiWords, 2013). The Harvard General Inquirer is a lexicon attaching syntactic, semantic, and pragmatic information to part-of-speech tagged words in a form of spreadsheet format (General Inquirer Home Page, 2002). LIWC, whose classifications are highly correlated with those of the Harvard General Inquirer, is a text analysis software program calculating the degree to which people use different categories of words across a wide array of texts, including emails, speeches, poems, or transcribed daily speech (What is LIWC?, 2014).

In order to determine the affective content of a post, sentiment scores are calculated for each post by means of SentiWordNet in the following way;

1. The word category for each word of a document is determined by a part of speech tagger.
2. For each ***adjective*** of a post, triple of polarity scores (positivity, negativity, objectivity) is computed by searching SentiWordNet. Since for one adjective, several entries may exist in SentiWordNet, the score of an adjective is calculated by

   $$score_{pol}\ A\ = \frac{1}{k}\sum_{k=0}^{n} score_{pol}(k) \tag{16}$$

   where *n* =number of synsets of term *A*, *pol* = pos, neg, obj, $score_{pol}\ A$ represents the positivity or negativity or objectivity score of term A and $score_{pol}(k)$ is a corresponding score of one SentiWordNet entry for this word. This results in a polarity triple for each sentiment-bearing term.
3. While computing a document polarity score triple, the three scores of each term are added individually. This sum of polarity score triples of all words is divided by the number of sentiment-bearing terms.
4. If $score_{pos}$ is larger than $score_{neg}$ the post is classified as **positive**. In case $score_{pos}$ is smaller than $score_{neg}$, it is considered **negative**; otherwise it is labeled **objective**.

Different classification algorithms have been tested in a 10-fold cross validation. Their data set consisted of over the 181 manually classified posts. For unsupervised clustering the *Density based clustering* performed best (accuracy of 73%), while for supervised classification the algorithms *Multinomial Naive Bayes* and *Logistic* achieved the best results (both with accuracy around 82%).

Denecke improved this approach in 2009 (Denecke, 2009). In this study, Denecke focused on applying the method in different domains, comparison of a rule-based and a machine-learning based approach for sentiment classification. The researcher analyzed accuracies achieved for different domains, and in cross-domain settings. The data set consists of documents of six different domains and two different types: product review, and news articles. 4872 negative and 5822 positive texts are available for the evaluation. The product review data set consists of Amazon product reviews. The reviews are available for four different product types: DVD, books, electronics and kitchen appliances. Each review consists of a rating (0-5 stars). Reviews with ratings larger than 3 were labeled *positive*, those with rating smaller than 3 were labeled *negative*, and the rest was discarded because their polarity was ambiguous. For each domain around 1000 *negative* and 1000 *positive* labeled examples are available.

Secondly, drug reviews from Drugratingz.com where users can anonymously rate drugs in several categories and can post comments were collected. Same assumptions were made and reviews with rating 3 were excluded.

A third data set was provided by the news articles given by the MPQA corpus (http://www.cs.pitt.edu/mpqa/) which consisted of news articles and was manually annotated with a variety of subjective information, such as subjective expression, objective speech event and agent.

In this paper, Denecke proposed 2 classification approaches; Rule-based Classification and Machine learning-based Classification. The rule-based classification approach first identifies opinionated words in a document based on SentiWordNet. If the positivity is larger, the word is considered *positive* and vice versa. If both values are equal, the word is ignored.

To determine the polarity of a complete document, the number of positive, negative and objective words is computed. Moreover, an average polarity score triple for the document is determined. If the number of positive words is larger than the number of negative words, the document is considered *positive* and vice versa. If the number of positive words equals the number of negative words, the average polarity score triple is checked and the document is considered *positive* when positivity value is larger than the negativity value. It is considered *negative* otherwise.

In the second approach, the machine-learning based classification first identifies 17 attributes for each text that are later used by a machine learning classifier:

Table 5: Feature sets for machine-learning based classification

Feature Sets	
Set 1	Average polarity score triples for adjectives, nouns and verbs (nine attributes),
Set 2	Frequency of positive and negative words (two attributes),
Set 3	Frequency of nouns, verbs, and adjectives (three attributes),
Set 4	Number of sentences, question marks and exclamation marks (three attributes).

For feature extraction, a text is first structured into tokens, stop words are removed and for each token, the absolute frequency and its part of speech are determined.

For evaluation purposes, Denecke measured the quality of the introduced approach in precision and recall. The results were tested for statistical significance using T-Tests. In the rule-based approach the accuracy values were reported very low between 40% for the classification of news articles to 65% for the classification of reviews of the electronic domain. This low accuracy was explained as due to the fix classification rule that could not handle different domain characteristics.

The machine-learning based approach was first studied for each domain separately in 10-fold-cross validations. Depending on the domain, the accuracy varies between 66% and 82%.

3.4 Classification based Quality Assessment Techniques

One of the most significant studies that lead the web content quality researches was handled in Discovery Challenge 2010 (DC2010) (The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) , 2010). Even if the domain was not health related and any kind of websites were assessed in the studies, the approaches presented in the competition was utilized in our study and consequently discussed in this section in detail.

The competition was sponsored by Google and administered by The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD).

The challenge focused on three different tasks;

(i) categorization of web contents (as spam, news, commercial, educational, discussion lists, personal, neutral, biased, trust and quality);
(ii) quality ranking of contents for English sites and
(iii) quality ranking for German and French sites.

In ECML/PKDD 2010 Discovery Challenge, the quality value is defined based on genre, trust, factuality and bias. Typically, DC2010 scores each host type empirically: The Spam host has quality 0; News/Editorial and Educational sites are worth 5; Discussion hosts are worth 4 while others are worth 3. Web spam, which gets the lowest score, is artificially-created pages that are injected into the web. Here the aim is to influence the results from search engines or to drive traffic to certain pages for fun or profit (Ntoulas, Najork, Manasse, & Fetterly, 2006). DC2010 also gives 2 bonus scores for Facts or Trust, but penalizes 2 scores for Bias hosts.

Participants were provided the data set that consisted of sample Web hosts from Europe with training and testing samples. No topic was specified. The features used in the competition were grouped under the two main classes; link based features and content based features. The link based features are the ones that provide information about the relationship between the host of the webpage and its neighbors and are highly utilized in the web spam detection algorithms. The features related to in-degree (the number of hosts that provide link to the host of the website) and out degree (the number of hosts that the website gives reference to) of the hosts and their neighbors, edge reciprocity (the number of links that are reciprocal), assortativity (the ratio between the degree of a particular page (number of inlinks or outlinks) and average degree of its neighbors) can be given as example. PageRank is a well-known link-based ranking algorithm which utilizes such kind of features to compute a score for each page. It is proposed by Larry Page in (Page & Brin, 1998). It is a mathematical algorithm based on the graph created by all World Wide Web pages as nodes and hyperlinks. The rank value indicates an importance of a particular page. The algorithm is based on the assumption that a hyperlink to a page is a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it (incoming links). A page that is linked to by many pages with high PageRank is given a high rank. On the other hand, if there are no links to a web page, it is assumed that there is no support for that page and a low rank is given.

Gyöngyi et al. proposed that if a page does not have any relationship with a set of known trusted pages, even if it has high PageRank, it is likely to be a spam page (Gyöngyi, Garcia-Molina, & Pedersen, 2004). TrustRank algorithm starts from a subset of predefined trusted nodes and propagates their labels through the graph. Although using this measure alone yields high number of false positives, it is very effective when used it with PageRank. Truncated PageRank is also derived from PageRank by Becchetti et al. (Becchetti, Castillo, Donato, Leonardi, & Baeza-Yates, 2006). In this algorithm, PageRank score of a page influences the score of its neighbors.

On the other hand, the content based features utilized in the challenge are the ones that are suggested by Ntoulas et al. to detect Web spam(Ntoulas, Najork, Manasse, & Fetterly, 2006). Ntoulas et al. generated the content based features by using the number of words on the page, the number of words in the title, the average word length, the fraction of anchor text (clickable text in a hyperlink which are used to determine the ranking that the page will receive by search engines), and the fraction of the visible text. Recall that web spammers can include unrelated text on the webpage by setting their visibility false to falsify the search engines.

In the challenge, Geng et al. was selected as the first place participant (Geng, Zhang, Jin, & Zhang, 2010). In their approach, Geng et al. utilized the statistical content features, page and host level link features given by the organization committee and the TFIDF features that they computed. They made use of information gain as the feature selection methodology which is known to be the best in text categorization, statistical spam filtering and information retrieval. They approached the problem as a classification problem in which the classes are 'Educational', 'Discussion', 'Commercial', 'Neutral', 'Bias' and 'Trusted'. They utilized the random decision tree methods C4.5. They reported that the classification accuracy has the highest score when all the features were included.

There are several issues that need attention in this study. They reported 0.936 accuracy in assessing the content quality which is quite high. However, they reported the accuracy in detecting the trustiness as 0.526 and bias as 0.606. This is because the content quality of a website does not have a clear cut definition. Consequently anyone can make his/her own assumption about the features of a qualified content and formulate the quality according to his/her own priorities. Here recall that the organization committee has formulated the content quality as an aggregate function of genre, (news-educational/discussion/commercial), trust, factuality and bias. The classification part has the highest weight in the assessment. As the classification is done, each website is assigned a score and the detection of trustiness or bias adds bonus or penalizes the score. The manual raters also evaluated the quality with the same methodology. As a result, the quality is mostly limited to the type of the website whether it is educational, commercial, or personal leisure which can be easily extracted from the domain name given in the URL.

3.5 Problems with the Current Techniques

Although both information relevance and quality are handled in the studies, these proposed methods are vulnerable to biased or even promotional contents which are created with commercial intent or amateurishly written(Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003). Such kind of biased contents will be ignored by the methods and the websites containing biased contents will get high scores if they contain the terms in the quality queries. Since the existence of the terms in quality queries within a website does not guarantee the absence of biased contents, these methods will be insufficient and will overestimate the quality scores. In their study Griffiths et al. have reported that they have selected their websites which were returned by human compiled search engines specific to depression (Griffiths, Tang, Hawking, & Christensen, Automated assessment of the quality of depression websites, 2005). Consequently it is unexpected that unprofessional contents were retrieved. Also the researchers expressed that they had excluded the websites that did not comprise core informational material and when there was no depression information on the primary site. In brief, biased contents were disregarded by Griffiths et al.

On the other hand, in the current temporal information retrieval techniques, although there are commonalities between our method and the other methods in the literature in terms of research goal, there are major differences. Firstly, in these studies, events are tried to be mapped to a timeline. Consequently the point is “when the events take place”. However, in our case, we focus on the research evidence about diabetes and analyze “when they are true or not”. In another words, our aim is to measure the extent to which the age of the data is appropriate. Secondly, approaches in temporal information retrieval do not assure quality of a retrieved document which will not work for health domain. However, by utilizing ADA guidelines we assure evidence based quality assessment in our framework.

While assessing the biased contents in health domain, detection of the commerce related synsets are important. Recall that information seekers are warned about the commercial contents while assessing the health related websites for quality (MedlinePlus Guide to Healthy Web Surfing, 2010). However, the negativity scores of the commerce related synsets are neutral in SentiWordNet. Consequently adaptation of the SentiWordNet for health domain is required to obtain efficient results.

It is also important to discuss the applicability of classification based quality assessment approach in the health domain. This approach is proposed to rank the websites in any domain. In this approach, a website gets 3 points when it is detected to be commercial and gets 2 points when it is detected to be trusted. However in the commercial websites, it is always challenging to understand if the authors have the intention to inform the consumers or they have only the commercial intents by displaying several advertorials on their web sites.

Another challenging issue is that there is a basic assumption that if a website is news website, it is highly possible for that website to be detected as of high quality which is again arguable for health domain. For instance, when it is news web page including facts, the content of the news will always be of high quality for years. However, contrary to the other websites, the health related websites may expire since the currency of information is nearly as important as the information itself in the health domain(Kim, Eng, Deering, & Maxfield, 1999).

In classification based quality assessment approach, information relevance is also disregarded and the websites are not ranked against their “information relevance”. In an effective search, data consumers expect the websites which are both relevant and qualified to be ranked at the top. Consequently, this approach could be valid only if it is assumed that the websites are already detected to be relevant.

CHAPTER 4

AUTOMATIC ASSESSMENT OF DIABETES WEBSITES FOR QUALITY AND BIAS

The previous chapters concluded that the studies in the literature are limited as they do not provide any ranking tool which automatically assesses quality concerning both bias information and timeliness aspects. In this chapter, our proposed methodology for quality assessment is introduced. This methodology aims to rank web pages according to information quality by taking into account both relevance and objectivity criteria.

4.1 The Method

The proposed methodology regarding the evaluation of bias and quality is based on sentiment analysis, relevance feedback and information retrieval techniques.

Figure 1 shows the overall framework. First, the crawled web pages are pre-processed for sentiment analysis and relevance feedback algorithms in the both training and testing data sets. The html tags are removed from the content. Second, the content is tokenized, and all upper-case letters are transformed into lower-case letters. Third, common words such as *a*, *an*, and *the* are removed. Since the biased websites significantly affect the information quality in a negative way, we identify such content by carrying out a *sentiment analysis*. Then, the quality queries are generated among the unbiased web pages in the training set, and they are utilized in the test set.

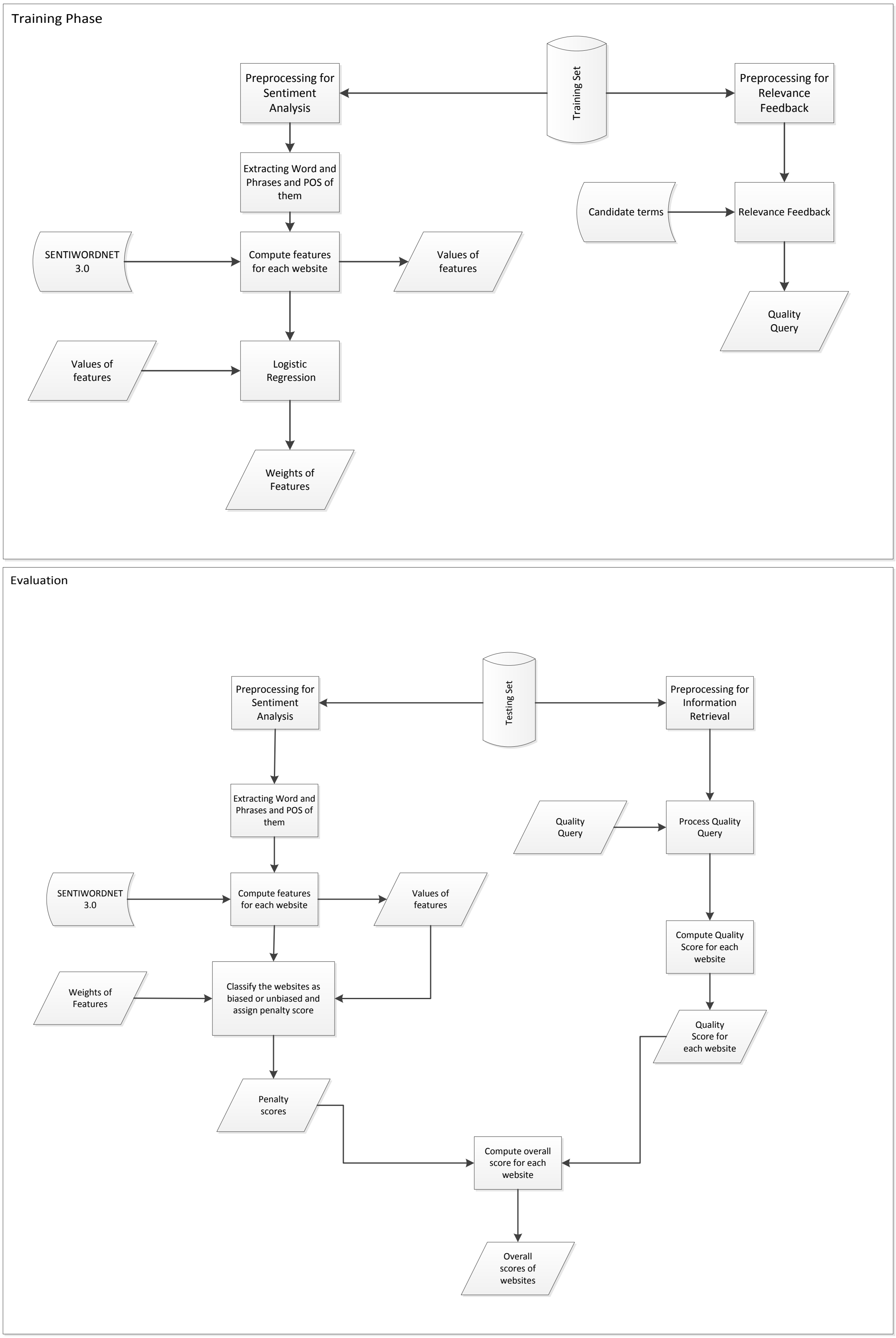
Training Phase
Preprocessing for Sentiment Analysis
Training Set
Preprocessing for Relevance Feedback
Extracting Word and Phrases and POS of them
Candidate terms
Relevance Feedback
SENTIWORDNET 3.0
Compute features for each website
Values of features
Quality Query
Values of features
Logistic Regression
Weights of Features
Evaluation
Preprocessing for Sentiment Analysis
Testing Set
Preprocessing for Information Retrieval
Extracting Word and Phrases and POS of them
Quality Query
Process Quality Query
SENTIWORDNET 3.0
Compute features for each website
Values of features
Compute Quality Score for each website
Weights of Features
Classify the websites as biased or unbiased and assign penalty score
Quality Score for each website
Penalty scores
Compute overall score for each website
Overall scores of websites

Figure 1: Overall Framework

4.1.1 Detection of Bias

The biased information in the health domain is defined as falsified contents that are written in deliberately obscure or unscientific sounding language, contain unrealistic health claims, and promise quick, dramatic, miraculous results (Griffiths, Tang, Hawking, & Christensen, Automated assessment of the quality of depression websites, 2005). U.S. National Library of Medicine National Institutes of Health warns information consumers against believing in web sites that claim that a specific remedy so-called "*a breakthrough medication*" will cure a variety of illnesses (National Institutes of Health's Web Site, 2012). It is also advised that non-commercial websites should be considered to be trustable resources, and if a drug is recommended by name, information seekers should check if the company that manufactures or sells the drug has provided that information.

In the proposed methodology, the bias information is identified with the help of the method proposed by Denecke (Denecke, 2009) but the method is improved for the research problem as follows: the negativity scores of the commerce related synsets are neutral in SentiWordNet. Since the purpose of the SentiWordNet is to aid in the identification of subjectivity but not the bias, both the positivity and negativity of the commerce related synsets such as *money*, *payment*, *shopping* or *credit card* are zero. In order to favor non-commercial content, the negativity scores of these synsets are assigned a value of 1. These terms are selected using the relevance feedback algorithm on biased websites. A total of 11 terms as shown in Table 6 were selected using this method. These terms were expanded to synsets with SentiWordNet.

Table 6: Commerce related terms

Terms
Bill
Buy
Sponsor
Charge
Shop
Pay
Money
Dollar
Credit Card
Price
Cost

In (Denecke, 2008), individual blog posts were classified by considering the presence of adjectives in the content. In our experiments, we aim to classify different types of websites about diabetes. However, this approach did not produce satisfactory results in terms of identifying the bias content as all the websites were labeled as objective. The ratios of the subjectivity score to the number of sentiment-bearing terms in large volume websites were very small.

As a consequence, the machine-learning based classification method proposed in (Denecke, 2009) is conducted. In each experiment, 17 attributes were identified to be used by a machine learning classifier. These attributes are average polarity score triples for adjectives, nouns and verbs (nine attributes), frequency of positive and negative words (two attributes), frequency of nouns, verbs, and adjectives (three attributes), the number of sentences, question marks and exclamation marks (three attributes). Then we compute the values of these features for each web page and used logistic regression in order to determine their

weights for differentiating biased web sites from unbiased sites (shown on the left of the image in
Figure 1 for the training phase).

4.1.2 Learning Quality Measuring Queries with Relevance Feedback

In the relevance feedback approach introduced in (Griffiths, Tang, Hawking, & Christensen, Automated assessment of the quality of depression websites, 2005), a complex query consisting of weighted words and phrases is automatically generated by comparing term frequency distributions in relevant and irrelevant documents. Here, the assumption is that terms in relevance query occur frequently in the relevant text but rarely otherwise. The resulting query is used by the text retrieval system to compute relevance scores for documents. In this study, this method is used to learn a *'quality'* query from sets of high and low quality webpages. So a quality query will comprise terms that will appear frequently in high quality web sites but rarely in low quality web sites. A website is categorized as high quality if it was given a score higher than 60 out of 100. Otherwise, it is accepted as low quality. The generated quality queries are then run on the websites to obtain quality scores for each webpage using the Terrier search engine, an open source search engine, (see the right side of the image in Figure 1 for the training phase).

In order to obtain a quality measuring query, firstly, the *candidate terms* for the quality measuring query are determined. These terms are obtained using:

- the list of words and phrases extracted from the list of quality criteria for diabetes websites given by (Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003).
- the top search terms for diabetes as can be seen in
- Table 7.

Table 7: Query term candidates gathered from top search terms

Diabetes Symptoms	Ketones
Gestational Diabetes	Alcohol and Diabetes
Type 1 Diabetes	Diabetic Neuropathy
Hyperglycemia	Pre-diabetes
Hypoglycemia	Diabetic Ketoacidosis

The frequency of each word and phrase is then computed for all the web pages in each website. Using these candidate terms and phrases, the weights are computed using the RSJ formula Equation 7.

The quality query terms are selected by computing the TSVs (Robertson S. , 1990) for each candidate term. The terms are ranked in descending order and the terms that had a rank above a certain threshold are selected.

4.1.3 Quality Scoring

In this step, the selected quality queries from the previous step are run using the Terrier search engine on the test data sets. The engine returns a score for each webpage and the score of a website is computed by calculating the average score of its web pages in accordance with the method proposed in (Griffiths, Tang, Hawking, & Christensen, Automated assessment of the quality of depression websites, 2005).

Okapi BM25, which was proposed by Robertson et al., (Robertson & Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, 1994) is used for the retrieval task which is introduced at Section 2.2.1. In the term frequency calculation, tf values tend to be larger in longer documents. Since it is more likely that a term will appear in longer documents, it causes a potential bias towards longer documents as they will be assigned higher scores. In order to compensate that effect, Robertson derived the document length normalization (Robertson & Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, 1994).

Okapi BM25 computes the score for each word or word phrase as given in Equation 5. The final score for a webpage is the sum of all the term weights as given in Equation 6. Finally, the tool returns a score for each webpage. The average of the score of individual web pages is computed to determine the final score of the web site.
In the training phase, the overall site scores are calculated as follows;

$$S_q = \alpha * \; Q_{score} + \; 1 - \alpha \; * Norm(\ln(|Q|)) \qquad (17)$$

Here, Q_{score} means the mean quality score obtained by Terrier and $Norm(ln(|Q|))$ is the normalized value of $ln(|Q|)$, obtained by dividing $ln(|Q|)$ by $ln(|Q_{max}|)$, where Q_{max} is the maximum number of retrieved documents per site.

In the training phase, we have the quality scores of each website (S_q) given by domain experts. As we calculate the $S^{BM25}(D,Q)$ by Terrier and $Norm(\ln(|Q|))$, the next step is to calculate an optimal α which maximizes the correlation between S_q and the right side of the equation. This parameter is used to adjust the balance between the average document score and the coverage of a site.

4.2 Dataset

Our data set consists of web sites collected and scored by Seidman et al., (Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003) and the bias web sites that we collected manually and were scored according to the same guidelines as given in (Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003) by a medical domain expert who is a pathology specialist. In order to select the bias websites, the criteria published by MedlinePlus (National Institutes of Health's Web Site, 2012) such as; sensational writing style, claims about remedies that cure a variety of illnesses or promise quick, miraculous results, and the existence of commercial advertisements were taken into consideration. We used the websites that were scored in the research carried out by Seidman et al. (Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003) since in their study the data set was constructed systematically taking into account various information quality problems which can also be considered as representative for thousands of real-world websites about diabetes. In their study, the websites were collected by querying a specific search term (i.e., diabetes) in Direct Hit search engine. The collected websites were manually scored between 0 and 100 by two domain experts based on the quality criteria prepared for Type 2 diabetes. The websites addressing only Type 1 diabetes or 'juvenile diabetes' were excluded due to the comprehensiveness criteria in the study (Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003). Also, the sites that only included news or did not offer general diabetes content were eliminated. At the end of these assessments, Seidman et al., scored 90 websites however in this thesis 41 websites were utilized since some of the links were broken, some were password protected, some were not allowed to be crawled or, only a single page from the site could be retrieved. The total number of web pages for each crawled website varied between 10 and 3500. Table 8 shows the list of the crawled websites. Some websites are diabetes specific (DB) and others provide general health information about several diseases so they were labeled generic (GN). In general health portals, Seidman et al., scored only the diabetes related webpages (Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003). In the experiments, since we aim to generate an overall information quality score for each website, we processed all the webpages from each web site. Since the quality and relevance of each web page might differ on a website, we scored each web page and calculated the average.

Table 8: The websites that were assessed in the study undertaken by Seidman et al., (Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003). We used the same web sites for our experiments. A total of 29767 web pages were processed for the experiments. Overall quality score implies the scores given for each web

	Links	**Overall Quality Score**	**Website Type**	**Number of Web pages**	**Size**
1	http://healthlink.mcw.edu/article	50	GN	1051	36 MB
2	http://my.webmd.com/index	63	GN	521	29 MB
3	http://www.banting.com/	59	DB	9	48 KB
4	http://www.bbc.co.uk/health/diabetes	50	GN	1138	63 MB
5	http://www.bddiabetes.com/	31	DB	387	9 MB
6	http://www.defeatdiabetes.org/	45	DB	271	8 MB
7	http://www.dhfs.state.wi.us/health /diabetes	78	GN	142	2 MB
8	http://www.diabetes.ca/	88	DB	1110	48 MB
9	http://www.diabetes.about.com	49	DB	510	12 MB
10	http://www.diabetes.org/	85	DB	757	59 MB
11	http://www.diabetesaustralia.com.au/	71	DB	21	488 KB
12	http://www.diabetesnet.com/	65	DB	547	40 MB
13	http://www.diabetesnews.com/	65	DB	99	3 MB
14	http://www.diabetesohio.org/	61	DB	63	3 MB
15	http://www.diabetic.org.uk/	73	DB	33	668 MB
16	http://www.docguide.com/	56	GN	2575	89 MB
17	http://www.dr-diabetes.com/	28	DB	9	152 KB
18	http://www.drkoop.com/	75	GN	2477	105 MB
19	http://www.drmirkin.com/diabetes	41	GN	122	1 MB
20	http://www.endocrineweb.com/ diabetes	33	GN	739	26 MB
21	http://www.evms.edu/diabetes	43	DB	521	12 MB

22	http://www.focusondiabetes.com/	65	DB	1	20 KB
23	http://www.healingwell.com/	53	GN	3185	88 MB
24	http://www.health.state.ut.us/cfhs	35	GN	41	780 KB
25	http://www.healthtalk.com/den/index	41	GN	4414	214 MB
26	http://www.idcpublishing.com/	41	DB	140	5 MB
27	http://www.idf.org/	47	DB	402	35 MB
28	http://www.joslin.harvard.edu/	73	DB	739	18 MB
29	http://www.lillydiabetes.com/	50	DB	31	1 MB
30	http://www.mayoclinic.com/	80	GN	217	9 MB
31	http://www.merck.com/pubs/ mmanual_home	56	GN	1333	71 MB
32	http://www.msdiabetes.org/	27	DB	165	3 MB
33	http://www.musc.edu/diabetes	18	GN	236	6 MB
34	http://www.netdoctor.co.uk/	73	GN	2067	96 MB
35	http://www.niddk.nih.gov/	65	GN	586	19 MB
36	http://www.nzgg.org.nz/library	42	GN	1349	73 MB
37	http://www.onlinemedinfo.com/	60	GN	744	25 MB
38	http://www.sddiabetes.net/	74	DB	4	88 KB
49	http://www.staff.ncl.ac.uk/philip.home	63	DB	6	380 KB
40	http://www.umassmed.edu/diabeteshandbook	68	GN	311	14 MB
41	http://uphs.upenn.edu/health	68	GN	694	20 MB

Table 9: The biased websites collected manually for the experiments.

	Links of Biased Websites	**Number of web pages**	**Size**
1	http://www.selfhelprecordings.com/diabetes/help-with-diabietes.asp	85	2 MB
2	http://www.your-diabetes.com/diabetes-supply.html	88	1 MB
3	http://shiningstarmiracles.wordpress.com/category/diabetes/	287	12 MB
4	http://www.holisticonline.com/Remedies/Diabetes/	606	13 MB
5	http://www.diabetes-daily-care.com/index.html	38	812 KB
6	http://prevent-diabetes.net/order.php	20	404 KB
7	http://www.diabetesdaily.com/forum/blogs/glodee/5936-incredibleendocrinologist	1022	52 MB
8	http://www.diabetes-supply.com/home.asp	16	612 KB
9	http://www.antioch.com.sg/well/testimon/muniandy/	1	12 KB
10	http://www.miraclesforyou.org	36	964 KB
11	http://christianblogs.christianet.com	1713	73 MB
12	http://www.d-mom.com/	358	13 MB
13	http://www .richardsearley.com	242	14 MB
14	http://www.hanselman.com/blog/HackingDiabetes.aspx	242	14 MB
	Total web pages processed	4754	

4.3 Experimental Settings

There are two objectives of the experiments:

1. The assessment of the performance of the proposed framework in identifying the biased websites
2. The assessment of the performance of the proposed framework in ranking the web sites according to their information quality and relevance

Table 10: Datasets used in the 5-fold cross validation

	Set 1	Set 2	Set 3	Set 4	Set 5	TOTAL
Biased Websites	3	3	3	3	2	14
Low Quality Websites	4	4	4	4	5	21
High Quality Websites	4	4	4	4	4	20
TOTAL	11	11	11	11	11	55

To achieve these objectives, 5-fold-cross validations are carried out since we had limited number of websites.

Table 10 gives the list of sets used in each experiment. In Table 11, the details of each set are given.

Table 11: The list of sets used in the experiments

Experiments	Training	Testing
Experiment 1	Set2 + Set3 + Set4 + Set5	Set1
Experiment 2	Set1 + Set3 + Set4 + Set5	Set2
Experiment 3	Set1 + Set2 + Set4 + Set5	Set3
Experiment 4	Set1 + Set2 + Set3 + Set5	Set4
Experiment 5	Set1 + Set2 + Set3 + Set4	Set5

The proposed methodology was executed separately for each experiment. We applied the processes under the training phase and evaluation phase in Figure 1 for the training and testing data set respectively. The training data set was used to determine the terms and their weights that are highly relevant with high quality web sites but are not relevant with low quality web sites. We used manually marked websites to construct these quality related terms.

Table 12 lists the terms and their weights computed by the RSJ formula in the five experiments. For example, Query 1 and Query 2 were formed using the training data set in Experiment 1 and Experiment 2 respectively in Table 11. In all the experiments, the terms that have a higher weight than 1.71 were selected and used in the remainder of the experiments. The threshold was selected by observing the cut-off point in the weight distribution.

Table 12: Terms and their weights. Terms, which were not extracted or had significantly lower weight than 1.71, are marked as "-".

Terms	Weights Query 1	Weights Query 2	Weights Query 3	Weights Query 4	Weights Query 5
blood_glucose_test	2.85	1.71	1.71	2.28	2.28
diabetes_prevention	2.85	2.28	2.28	2.69	2.69
treatment_type	2.85	2.28	1.71	1.71	2.85
treatment_type_diabetes	2.85	2.28	1.71	1.71	2.85

healthy_eating	1.88	1.89	1.89	3.56	2.43
obesity	1.88	3.02	-	1.89	2.17
acromegaly	1.71	1.71	1.71	-	1.71
diabetes_obesity	1.71	1.71	-	1.71	1.71
diabetic_complications	1.71	1.71	-	1.71	1.71
eating_disorders	1.71	1.71	1.71	-	1.71
family_history	1.71	1.71	-	1.71	1.71
insulin_administration	1.71	1.71	1.71	1.71	-
meal_planning	1.71	2.28	2.69	2.69	2.69
diabetes_pregnancy	-	1.71	-	-	-
eating_healthy	-	1.71	1.71	1.71	1.71
insulin_dependent	-	1.71	-	-	-
meal_planning guide	-	1.71	1.71	1.71	1.71
pre-diabetes	-	1.71	-	-	-
medications	-	-	3.02	-	-

The terms were later utilized in the testing data set to calculate a quality score. We then calculated the true positives, false positives and accuracy for each experiment. We also utilized Pearson correlation to measure the performance of the proposed framework's ranking with manually scored ones by the domain experts. If the correlation is high and positive, it will indicate that the ranking is in line with the ground truth information. In the training phases, all the websites given by Seidman et al., (Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003) (in Table 8) were labeled as unbiased since none of them contained biased content.

To achieve the optimal value for α, all the values between 0 and 1 were tried with an increment of 0.01. The value which maximized the correlation between the computed site scores, Q_{score}, and manually assigned score S_q was chosen in each training set. In five experiments, the maximum correlations were achieved when α value was nearly 0.8.

4.4 Tools

RapidMiner is a well-known widely used open source system for data mining, text mining and predictive analytics (RapidMiner, 2014). We chose this tool for this study since it is stable and powerful due to its flexibility in the process design (Data Mining Tools Used Poll, 2013). Terrier, an open source search engine, was used for information retrieval tasks (Terrier IR Platform v3.5, 2014). We used the Porter Stemmer algorithm (Porter, 2006) a commonly known and widely used stemmers for English language words and it is supported by RapidMiner (Willett, The Porter stemming algorithm: then and now, 2006). The part of speech (POS) of each word in a sentence was identified by TreeTagger (Schmid, 1994).

4.5 Results

In this section, the performance of the proposed framework in terms of accuracy in biased website detection and quality scoring will be presented.

For bias content detection, the algorithm produced the results as shown in Table 13. In the evaluation, five sets of experiments were performed each containing 11 websites. Of 14 biased websites, 11 websites were detected correctly as biased whereas 3 unbiased websites were misclassified as biased. Consequently, a total of 6 out of 55 websites were misclassified; giving an accuracy of 89%.

Table 13: Results in bias content detection

Experiments	**Testing**	**Number of Biased Web Sites**	**Number of Websites Detected as Biased**		
			True Positive	**False Positive**	**Accuracy**
Experiment 1	Set 1	3	2	1	0.82
Experiment 2	Set 2	3	2	1	0.82
Experiment 3	Set 3	3	3	0	1.00
Experiment 4	Set 4	3	3	1	0.91
Experiment 5	Set 5	2	1	0	0.91
Total		14	11	3	0.89

In the detection of bias content, the best results were obtained in the Experiment 3 in which all the biased contents were detected, and no unbiased contents were misclassified. The worst results were retrieved in Experiments 1 and 2 in which the accuracy was 0.82.

The biased websites, which were not detected by the proposed system obtained high scores in the quality scoring since no penalty score was applied. So the correlation between the computed and actual scores decreased dramatically.

For evaluation purposes, the Pearson correlation between the quality scores given by the domain experts and the scores generated by the proposed framework were calculated and compared with the other techniques. As shown in Table 14, the first row shows the results of the proposed framework. The second row gives the results of the model proposed by Griffiths et al. (Griffiths, Tang, Hawking, & Christensen, Automated assessment of the quality of depression websites, 2005) and the results clearly illustrate the limitation of their method. In the last row, the results show the limitation of using only the keywords and weights of SentiWordNet in bias website detection. When we did not modify the weights of the proposed terms as presented in Table 14, a lower ρ value was obtained. The scatter plot in Figure 2 illustrates the degree of correlation between these variables and suggests that the correlation is positive.

Table 14: The results of the experiments

	Experiment 1	**Experiment 2**	**Experiment 3**	**Experiment 4**	**Experiment 5**
ρ (p<.001) (proposed framework)	**0.59**	**0.62**	**0.73**	**0.82**	**0.62**
Griffiths et al. technique (Griffiths, Tang, Hawking, & Christensen, Automated assessment of the quality of depression websites, 2005)	0.38	0.26	0.32	0.36	0.35
ρ with	0.49	0.36	0.33	0.81	0.41

original SentiWordNet					

In 5 test sets, the best results were obtained in Experiment 4 in which the correlation was 0.82. In each test set the correlations were significantly higher in the proposed framework.

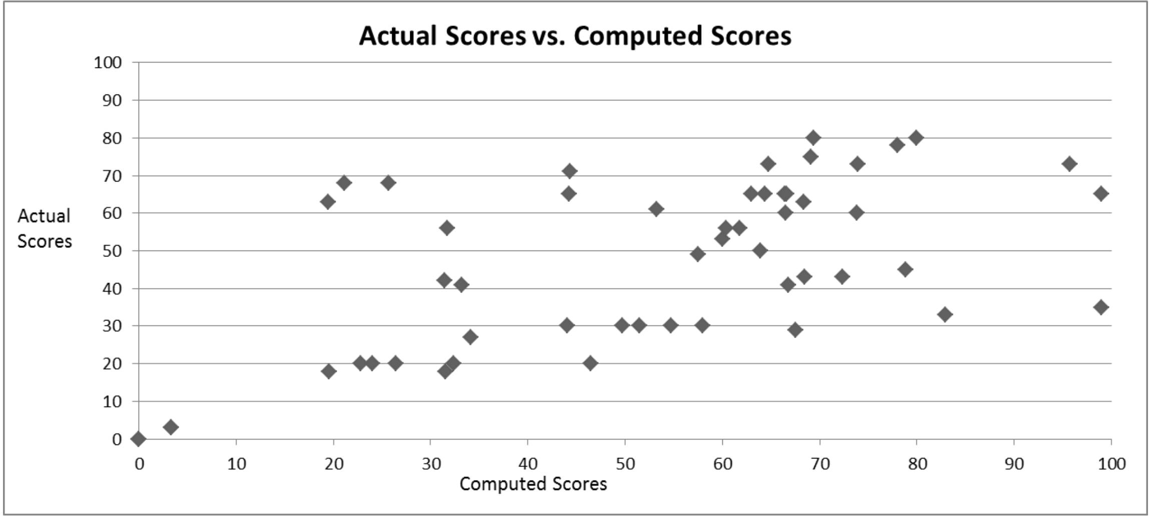

Figure 2: The correlation between the actual scores and the computed scores for the proposed framework within 5 sets.

In order to evaluate the effect of the proposed changes made to SentiWordNet, an attempt was made to identify the biased content using the original SentiWordNet. As a result, we obtained lower accuracy values compared to the proposed method as shown in Table 15. The reason for this is due to the fact that since the commercial content provides one of the most important pieces of evidence for the identification of the bias content in the health domain, some of the weights of the terms need to be penalized in SentiWordNet in order to disfavor commercial content. The Pearson correlation between the actual scores and the computed scores based on the original SentiWordNet is decreased compared to the proposed framework as given in

Table 13.

Table 15: Results in bias content detection with original SentiWordNet

Experiments	Testing	Number of Biased Web Sites	Number of Websites Detected as Biased		
			True Positive	False Positive	Accuracy
Experiment 1	Set 1	3	2	1	0.82
Experiment 2	Set 2	3	1	1	0.73
Experiment 3	Set 3	3	1	3	0.55
Experiment 4	Set 4	3	3	0	1.00
Experiment 5	Set 5	2	0	1	0.70
Total		14	7	6	0.76

The low correlations between the actual and computed scores in certain test sets necessitated further analysis to present a detailed explanation. The main characteristic of the websites in Experiment 1 was that most of the websites in this set were not diabetes specific websites and provide general health content about several issues. For example *bbc.co.uk*, *webmd.com*, *dhfs.state.wi.us* and *healthlink.mcw.edu* are in this set. Consequently, the scores of many unrelated webpages about diabetes were included in the computation which averages all the scores of each website page even there was only a link to diabetes related web page. In order to present the effect of website characteristics, a new experiment was designed. The websites were placed in 2 groups; general purpose websites and diabetes specific and then 16 websites were randomly selected from each group. The 5 biased websites were added to both test sets, and the quality query 4 (in

Table 12) was run against the collections and the Pearson correlation was computed between the computed and actual scores.

Table 16: General websites and diabetes specific websites

	General Websites	**Diabetes Specific Websites**
Biased Websites	5	5
Low Quality Websites	8	8
High Quality Websites	8	8
TOTAL	21	21

A lower correlation of 0.25 was observed in the test set containing general websites. In large volumes websites, Terrier retrieved more than 1000 pages most of which had very low scores that were close to zero. The low scored webpages were those in which diabetes was not the main concern but there were some links to diabetes related webpages. Since the links contained diabetes specific terms, they were retrieved by Terrier even if with low scores. The low scores lowered the average, and all the websites were scored between 30 and 40 over 100 by the system.

In diabetes specific websites, the low scores close to zero were still retrieved however, they were fewer in number. These pages were generally ‘About us’ or ‘Contact us’ pages. The correlation was computed as 0.76 (p<.001). Both results are given in

Table 17.

Table 17: Results from general websites and diabetes specific websites

	General Websites	**Diabetes Specific Websites**
ρ $(p<0.001)$	0.25	0.76

4.6 Discussion

Assessing the information quality of health web sites automatically is challenging since it is necessary to simultaneously take into consideration many issues such as accuracy, bias, information relevance, and timeliness.

In this thesis, we propose a framework which aims to provide a better identification and ranking of diabetes web sites according to EBM. Previous approaches in the literature are either manual or limited to addressing wide-ranging information quality problems. The results showed that the proposed framework had a significantly higher ρ value compared to the other techniques (the average of all the experiments was 0.68 compared to the other technique where an average ρ of 0.33 was obtained). It also identified the bias websites with high accuracy. A high correlation between the manual ranking that was carried out using EBM by the domain experts and our proposed framework suggests that the method is able to generate a successful ranking according to EBM.

Although the results of the study are promising, there are still some limitations. First, by calculating the average scores of the individual webpages of the web sites, health portals that are not diabetes specific are automatically penalized in this method.

Assigning a static penalty score is the second deficiency of the study. Dynamic penalty scores should be applied to the websites in relation to its biased content ratio over the whole content. Although the current approach affected the results in a positive way, the linear static combination is not a desired solution.

CHAPTER 5

AUTOMATIC TIMELINESS ANALYSIS OF DIABETES WEBSITES

The part of the framework introduced at Chapter 3 describes the steps that differentiate high quality and unbiased contents from the low quality or biased ones. However it ignores the variety of the high quality contents regarding the information timeliness aspects. It assigns similar high scores to websites that were published in 2011 and 2013 with the information available at these periods. However, EBM approach requires websites to accommodate their contents as the new guidelines published each year.

On the other hand, as concluded in Chapter 3, even there are studies regarding the timeliness evaluation of contents, the requirements for health related websites are different. Studies based on temporal language models in the literature are generally utilized for mapping events to a timeline (Kanhabua & Norvag, Using temporal language models for document dating, 2009). Events occur at certain periods of time which has certain start and end dates. Here, we focus on facts about diabetes all of which have been introduced to the literature at a certain date and they become valid for a long time (majority has no end date). However in health domain, the requirement is different from other existing studies in temporal information retrieval domain which focus on extraction of time related information (temporal entities) from content to predict the exact time the web page belongs to (Alonso, Gertz, & Baeza-Yates, 2009) (Klimov, Shahar, & Taieb-Maimon, Intelligent visualization and exploration of time-oriented data of multiple patients, 2010). Such approaches are not suitable in this domain since although many dates are updated on the web pages, content may not reflect all up-to-date information in health domain. Consequently, rather than using temporal expressions, entire document content should be utilized to assess timeliness.

In this chapter, the part of the framework which automatically assesses the content of health web sites and predicts to which time period a given content belongs to regarding the evidence based medicine is introduced. To the best of our knowledge, this study is the first of its kind in the literature.

5.1 The Method

The method comprises three main steps: Term recognition, query generation and web site scoring

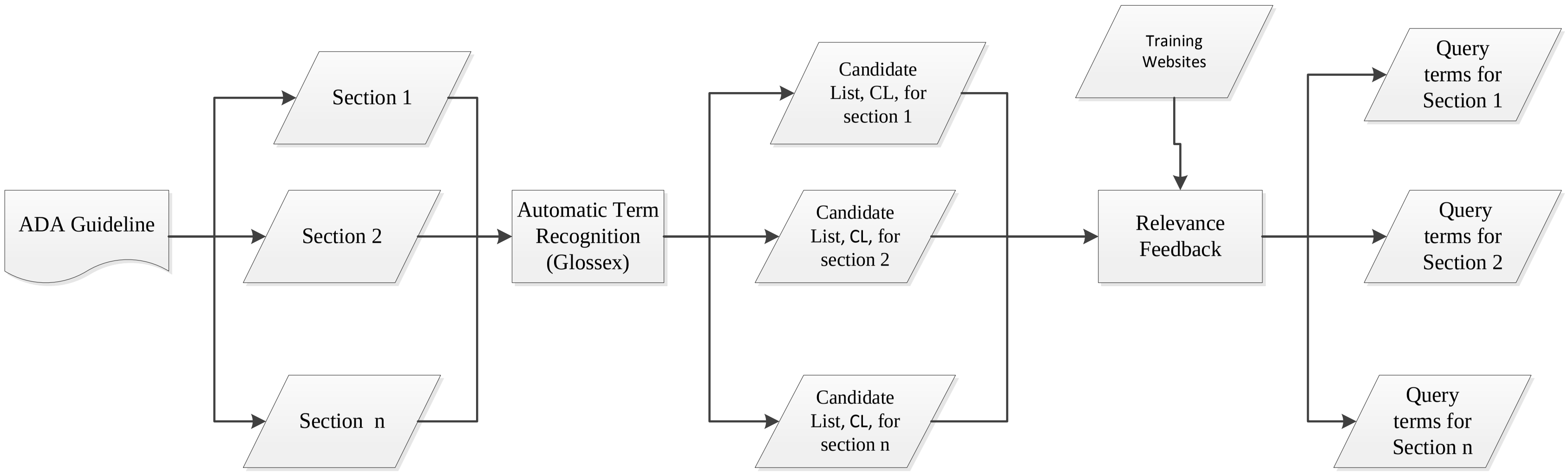
ADA Guideline
Section 1
Section 2
Section n
Automatic Term Recognition (Glossex)
Candidate List, CL, for section 1
Candidate List, CL, for section 2
Candidate List, CL, for section n
Training Websites
Relevance Feedback
Query terms for Section 1
Query terms for Section 2
Query terms for Section n

Figure 3: The flowchart of the training phase. This process is carried out for each publication year of ADA guideline.

5.1.1 Step 1: Term Recognition

Input: Term recognition is carried out on ADA guidelines (the selected reference guidelines (*RG)* for diabetes domain) archives of which are retrieved from their web sites (American Diabetes Association, 2011). The guidelines published between 2006 and 2013 are downloaded and all the sections of the guidelines are covered.

Method: As summarized in Figure 4, the terms are extracted from each ADA section to obtain candidate lists which resulted in 35 term lists using ATR techniques. In ATR techniques, the aim is to extract words and multi-word expressions that are significant for a given domain(Knoth, Schmidt, Smrz, & Zdrahal, Towards a framework for comparing automatic term recognition methods, 2009). ATR methods in the literature are generally used for keyword extraction and ontology enrichment (Knoth, Schmidt, Smrz, & Zdrahal, Towards a framework for comparing automatic term recognition methods, 2009). In this study, this approach is used to extract candidate terms from each section of the ADA guideline. Although Glossex is selected as the base method in this step, other well-known ATR models described at Section 2.2.2 are also applied to compare their effects in capturing significant terms in the experiment section.

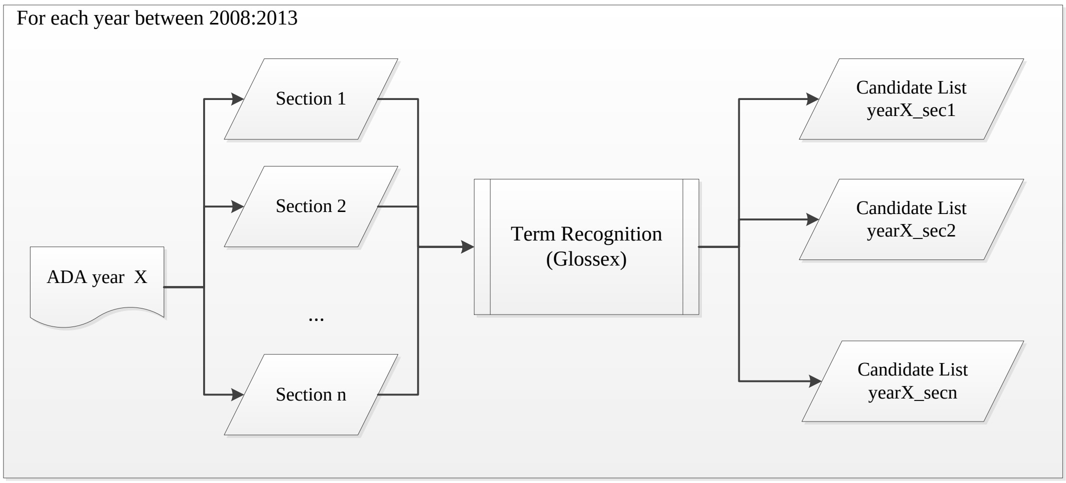
For each year between 2008:2013
ADA year X
Section 1
Section 2
...
Section n
Term Recognition (Glossex)
Candidate List yearX_sec1
Candidate List yearX_sec2
Candidate List yearX_secn

Figure 4: Term Recognition Phase

Output: At the end of this phase, the list of terms and their ATR scores for Glossex, Likelihood and C-Value methods are generated for comparison purposes. Candidate lists (*CL*) are created based on the Glossex scores. The reason to apply Glossex for term recognition is discussed in Section 5.5. Section specific *CL*'s are generated for each guideline.

5.1.2 Step 2: Query Generation

Input: *CL* generated at Step 1 and training dataset comprising both high quality websites, D_H, and low quality websites D_L are inputs in this step.

Method: The extracted terms in *CL* in Step 1 are used to generate section specific queries to be utilized in modelling of the evolution of each ADA section within the last six years. As given in Figure 5 the queries are produced using Relevance Feedback as done at Section 4.1.2. Recall that relevance feedback is an automatic process for query reformulation by choosing important terms or expressions attached to certain previously retrieved relevant documents and enhancing the importance of these terms in the final query(Griffiths, Tang, Hawking, & Christensen, Automated assessment of the quality of depression websites, 2005). In our case, term frequency distributions in D_H and D_L are compared and are used to generate a complex query consisting of weighted words and phrases. The terms that appear on both D_H and D_L are given low weights whereas the ones that appear only on D_H are assigned to higher weights. The weights in the queries are computed using the RSJ formula given at Equation 7. This formula is computed for each query term $qt_{ik}(y)$ (k^{th} query term in *CL* of the i^{th} section of the selected *RG*) in each year *y*.

RSJ assigns high scores to the terms that appear mostly in highly relevant web pages(Robertson & Jones, Relevance weighting of search terms, 1976). The query terms $qt_{ik}(y)$ that scored greater than 0 by RSJ are selected (Robertson S. , 1990). Then, the terms $qt_{ik}(y)'$ are ranked in descending order and the terms of which have a rank above a certain threshold th_{iy} (a cut off point for section *i* and year *y*) are chosen.

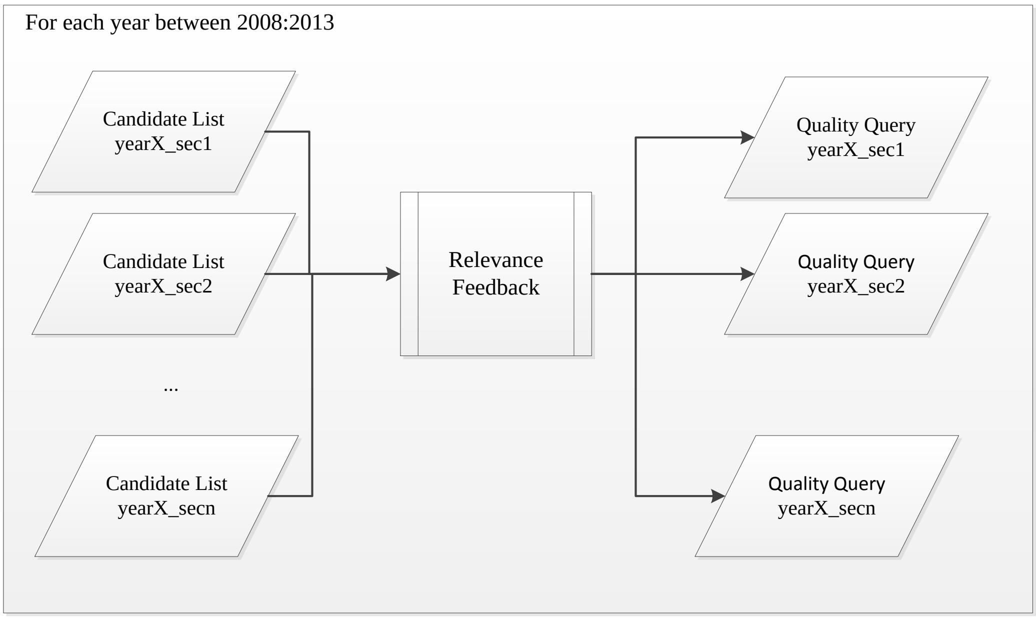
For each year between 2008:2013
Candidate List yearX_sec1
Candidate List yearX_sec2
...
Candidate List yearX_secn
Relevance Feedback
Quality Query yearX_sec1
Quality Query yearX_sec2
Quality Query yearX_secn

Figure 5: Query Generation Phase

Output: Query generation phase results in several queries which comprise terms $qt_{ik}(y)'$.

5.1.3 Step 3: Web Site Scoring

Input: The test dataset *D′* which includes high quality websites with actual update time labels and the generated queries at Step 2 are the inputs of this phase.

Method: The generated queries in Step 2 are used by the text retrieval system to compute scores for *D′* and the score of each website is computed by calculating the average score of its web pages(Griffiths, Tang, Hawking, & Christensen, Automated assessment of the quality of depression websites, 2005). Okapi BM25 (Equation 5) is utilized for the retrieval task (Robertson & Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, 1994). Recall that in this formula, *tf* (term frequency) and i*df* (inverse document frequency) values and the relative length of a document are taken into account. Okapi BM25 computes the score for each term t Є $qt_{ik}(y)'$ in *D′* in the query. Note that this process is undertaken for all years and all sections of *RG*.

The final score for a web page, S_{iy}^{p}, is the sum of all the term weights as given in Equation 6. The score of a website S_{iy}^{w} is computed by calculating the average score of its web pages S_{iy}^{p} where *i*, *y*, *p* and *w* imply section, year, page and website in *D*' respectively.

For timeliness analysis, a website is scored against each *RG* section (in this case ADA guideline's sections) for each year between 2008 and 2013. Section specific queries *qt*$_{ik}$*(y)'* generated in Step 2 are run and the query which returns the highest score S_{iy}^{w} is used to determine the update time of the website. This process is repeated for each section i and the update time of the website is predicted as the most common year assigned by all sections. The process is summarized at
Figure 6.

The formulas are given below. For the following equations, w is the website, y is the year that the context belongs to, and S_{iy}^{w} is the quality score obtained for the year y and for the section i.

$$Update\ time_i^w = argmax_{\ y}\ (S_{iy}^{w}) \tag{18}$$

$$Update\ time^w = mode(Update\ time_i^w) \tag{19}$$

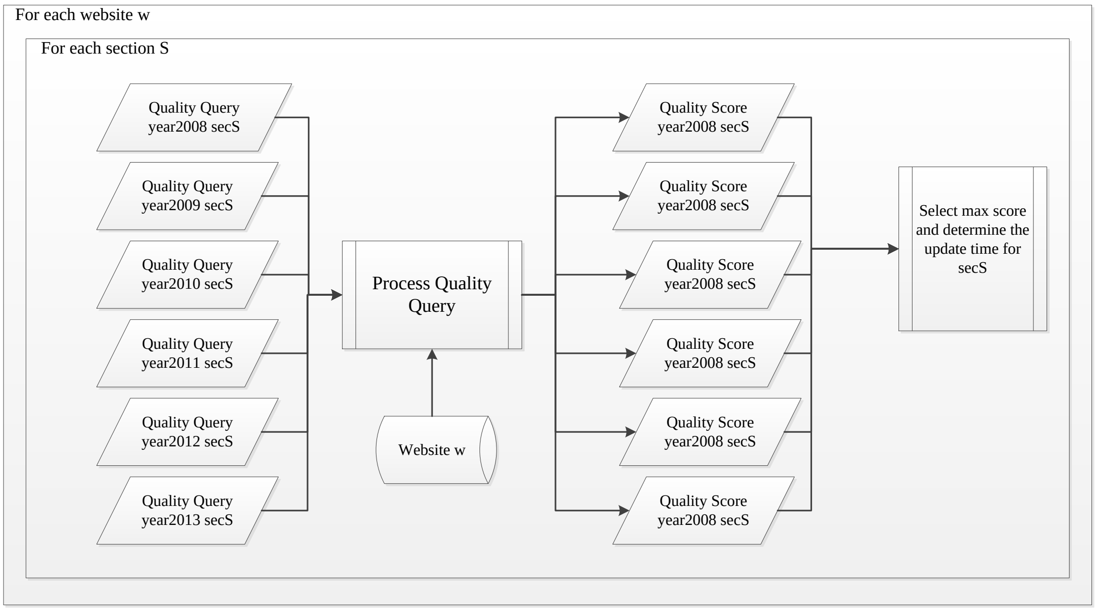
For each website w
For each section S
Quality Query year2008 secS
Quality Query year2009 secS
Quality Query year2010 secS
Quality Query year2011 secS
Quality Query year2012 secS
Quality Query year2013 secS
Process Quality Query
Website w
Quality Score year2008 secS
Quality Score year2008 secS
Quality Score year2008 secS
Quality Score year2008 secS
Quality Score year2008 secS
Quality Score year2008 secS
Select max score and determine the update time for secS

Figure 6: Web Site Scoring Phase

Output: Section specific website scores provide valuable information with respect to the timeliness of a given web site according to different topics regarding diabetes since timeliness is computed using each *RG* section. On the other hand, the update time of a web site $Update\ time^w$ gives an overall idea the timeliness of the specific website.

5.2 Dataset

As there is no standard data set to test the proposed method in the literature, we have constructed the data set. As a trusted and up-to-date resource to be used in order to measure the currency of web sites, ADA guidelines which were published between 2008 and 2013 are selected.

All the archives of ADA guidelines are retrieved from their web sites (American Diabetes Association, 2011) which are publicly available. The web sites are selected by querying specific search terms “diabetes” and “diabetes mellitus” in HON-search which returns the websites subscribed to the Code of Conduct (HONcode) principles (HON, Health On the Net Foundation, 2010). Recall that, as introduced at Section 2.1.9, The HONCode is the oldest and the best known quality label on the Web developed by a non-profit organization HON which works under the auspices of the Geneva Ministry of Health in Geneva, Switzerland.

In the data collection step, the initial HON review and subsequent monitoring dates are paid attention in order to select the corresponding archive copy of the web sites from Archive-it (Archive-It: A service of the Internet Archive, 2013). Archive-it is a subscription web archiving service from the Internet Archive (Internet Archive, 2001) that helps organizations to harvest, build, and preserve collections of digital content. Since Archive-it does not always archive all the websites regularly, those that are missing could not be retrieved.

This study is mainly focused on the timeliness analysis of high quality websites since low quality websites already suffers from many other significant factors such as misinformation or inaccurate information so timeliness cannot be carried out for these web sites properly. However, low quality web sites given at Table 21 (the last three rows: bddiabetes.com, cdc.gov and idf.org) are also crawled since the proposed method requires low quality web sites in the training phase. However, they are not included in the testing phase.

These low quality web sites are selected from the study of Seidman et al. (Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003) and are crawled on Archive-it. Recall that Seidman et al. collected the websites by querying a specific search term (i.e., "diabetes") in Direct Hit search engine and the medical experts scored them between 0 and 100. In order to evaluate the websites, the researchers created a data abstraction tool in a form of questionnaire which included comprehensive evaluation criteria such as; required explanations, the validity of methods used, currency and accuracy of information.

In this study, bddiabetes.com, cdc.gov and idf.org web sites are utilized which are scored as 31, 35 and 47 respectively in the study of Seidman et al.(Seidman, Steinwachs, & Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites, 2003). The two captures that were published in 2006 and 2007 are crawled for each website. As a result, 46 captures of 118 websites are obtained in the study.

5.3 Experimental Settings

The reference document (ADA guideline) and the input web pages are pre-processed before conducting the analyses. They are tokenized and the stop words are filtered. Although ADA guidelines are published each year, there may be no significant differences between some of its sections in subsequent years. For example, the same nutrition recommendations appeared both in 2007 and 2008. In 2007, Referral for Diabetes Management section was published as it was in 2006. These sections are ignored in this study since they do not provide valuable information for timeliness analysis. On the other hand, major revisions took place in some sections of the document such as Detection of GDM in 2007, Hypoglycemia in 2008 or Pharmacologic and Overall Approaches to Treatment in 2013. As a consequence, while some sections appear as they are in several years, some others are updated with major revisions each year. Due to this fact, rather than aiming to estimate the timeliness of the web site content in years, this study targets to predict the timeliness of the web site content in non-overlapping consecutive time intervals (i.e. 2006-2008, 2008-2010). To determine the time intervals, the content variations in the guidelines are analyzed in detail manually and the findings are confirmed by cluster analysis results. There are 35 different sections in the guidelines published so far since 2006.

The variation in the guidelines suggests three time intervals. To confirm this finding, documents are also clustered using k-means algorithm. In k-means clustering, each object is assigned to precisely one of a set of clusters and the similarity between objects is calculated using the squared Euclidian distance between them in a maximum of 10 runs. The clustering results also showed that there are 3 clusters.

Table 18: The clustering results of ADA guidelines published between 2005 and 2013

Cluster No	Years
Cluster-1	2005, 2006, 2007, 2008
Cluster-2	2009, 2010, 2011
Cluster-3	2012, 2013

As seen in Table 18, the clusters produced by k-means algorithm comprise the guidelines published in successive years. The results suggest that the document revisions between 2008-2009 and 2011-2012 are substantial but the revisions between 2005- 2008, 2009-2011 and 2012-2012 are insignificant.

In addition, the ADA document should not be considered as a single entity to estimate the timeliness of a given document. ADA guideline's section descriptions vary in length significantly although each section has equal importance. While applying text processing methods in the whole document, the impact of the longer sections might be higher compared to the smaller sections. Thus, each section of the ADA document is treated separately in this study. In other words, text processing methods are applied to each individual section separately. Table 26 in Appendix also shows the evolution of the guidelines considering the number of words used in the sections within 8 years.

The experiments comprises 3 main steps; automatic term recognition, timeliness prediction according to years and timeliness prediction according to time intervals. Term recognition experiments are accomplished during training phase and timeliness predictions are undertaken under the testing phase.

5.4 Tools

In this study, for data mining tasks such as k-means clustering, RapidMiner is used (RapidMiner, 2014). The information retrieval tasks such as query generation are conducted using Terrier Information Retrieval Platform 3.5 (Terrier IR Platform v3.5, 2014). We also use the Porter Stemmer algorithm (Porter, 2006) a commonly known and widely used stemmer for English language words (Willett, The Porter stemming algorithm: then and now, 2006) which is also supported by RapidMiner.

5.5 Automatic Term Recognition

The training phase in the experiments starts with term recognition. Recall that Glossex Method has been reported as a superior method in term extraction in the literature(Zhang, Iria, Brewster, & Ciravegna, A comparative evaluation of term recognition algorithms, 2008). Due to this reason, we utilize Glossex to construct the term lists. However, to confirm its advantage, an experiment is also conducted in order to measure whether it is able to capture the relevant and significant terms from the guideline. For the methods that require background knowledge, the Open American National Corpus (OANC) is used as a general corpus which includes 14.6 million words (American National Corpus, 2009).

As an evaluation metric, the precision of the methods are reported at 3 points (cuts); first 20 highly ranked terms, first 200 and first 2000 terms which are widely used in the literature for the evaluation of ATR methods(Knoth, Schmidt, Smrz, & Zdrahal, Towards a framework for comparing automatic term recognition methods, 2009). The precision is defined as;

$$Precision = \frac{{}_{i=0}^{|Recognized|} |t_i \in Reference|}{|Recognized|} \qquad (20)$$

where $Recognized$ is a set of highly ranked terms extracted by the method and $|t_i \in Reference|$ is 1 if term t_i is in the $Reference$ set containing the list of correct terms. Otherwise it is set to 0. In this experiment, two $Reference$ sets are manually generated to be able to evaluate the metrics: the first list comprises the terms that are added to the guidelines for the first time and the second list includes all the terms in the guidelines. The first list is important since for timeliness analysis, the newly added terms are critical elements. For example all the diabetes web pages are expected to include *diabetes* term but not a term such as VEGF therapy which was first introduced in the literature in 2013. Table 19 shows the list of terms that are added to the guideline for the first time within the last 8 years.

The experiment confirms the findings in the literature for ADA guidelines (Knoth, Schmidt, Smrz, & Zdrahal, Towards a framework for comparing automatic term recognition methods, 2009) Table 20 shows the precision results of Weirdness, Glossex, Likelihood Ratio and C-Value.

Table 19: The newly added terms in ADA guidelines

Year	**Terms**
2006	HbA1c, Nephropathy, Protein restriction/intake, Chronic kidney disease (CKD), Serum creatinine
2007	Emergency and disaster preparedness, waterproof and insulated disaster kit, glucose testing strips, glucose-testing meter, cool bag, antibiotic ointments/creams, glucagon emergency kits, Celiac disease
2008	Hypoglycemia unawareness, Hypothyroidism, Thyroid peroxidase, Thyroglobulin antibodies, TSH, Thyroid dysfunction, Thyromegaly, Intravenous insulin protocol
2009	Bariatric surgery, Pneumococcal polysaccharide vaccine, Foot pulses, Pinprick sensation, Ankle reflexes, Vibration perception threshold, Intravenous insulin infusion
2010	Fundus photographs
2011	Biopsy, EGFR, Estimated gfr, fasting plasma glucose, frequent unexplained hypoglycemia, Diarrhea, malabsorption
2012	Driving and Diabetes, Comorbidities of diabetes, Pharmacotherapy for hyperglycemia, Camps
2013	Cognitive function, Hepatitis B, Urinary albumin excretion, Anti–vascular endothelial, growth factor, VEGF therapy

Table 20: The precision results of the methods. P@20, P@20 and P@200 refer to precision at20, 200 and 2000 respectively based on the second Reference Set and P@T indicates precision at detecting new terms based on the first Reference Set The best results are obtained using Glossex method which ranked first in performance in 3 out of 4 cases.

Weirdness and Glossex produce similar results which can be explained as Glossex is the extension of Weirdness.

	Weirdness	Glossex	Likelihood Ratio	C-Value
P@20	0.90	0.90	1	1
P@200	0.93	0.94	0.94	0.85
P@2000	0.86	0.87	0.46	0.77
P@T	0.93	0.94	0.22	0.90

5.6 Timeliness Prediction According to Years

The queries in Step 2 of the training phase are generated using the training data set in Table 21. The first two web sites are the high quality web sites whereas the latter three web sites are low quality web sites (so they do not have any HON Monitoring Dates).

The training data set comprises 2 high quality web sites' (*D_H)* archives and 3 low quality web sites' (*D_L)* archives. For low quality websites archives, the captures date of which are close to 2006 are preferred. The captures of two belong to 2006 and 2007 since for some web sites, 2006 version was not archived by Archive-it. For high quality websites of each year, their corresponding archives in Archive-it are utilized.

The same D_L are utilized in Equation 7 for all years since our focus is to measure the timeliness aspects of D_H. Table 22 shows the queries $qt_{ik}(y)'$ generated specific to section "Nephropathy Screening and Treatment" in *RG* between 2008 and 2013. The values in the table are the weights of the terms in the corresponding queries.

Table 21: Training Dataset

Website	**URL**	**HON Monitoring Dates**	**Capture Dates of Archive-it**	**Website Features**		
				Year	**Number of pages**	**Size**
American Diabetes Association	http://www.diabetes.org/	May 29 2010, Mar 24 2010, Mar 09 2010, Feb 22 2010, Dec 11 2009, Nov 26 2009, Mar 03 2006, Feb 24 2006, Apr 17 2003, Jul 30 1997	Oct 2008, Apr 2009, Apr 2010, June 2011, Jun 2012, Jun 2013	2008	1492	141 MB
				2009	849	79 MB
				2010	574	54 MB
				2011	572	54 MB
				2012	421	35 MB
				2013	199	20 MB
Indian Health Services	http://www.ihs.gov/	Aug 21 2013, Feb 28 2011, Feb 10 2011, Jan 26 2011, Jan 27 2010, Jan 12 2010, Jul 22 2008, Jan 27 2008	2006, Feb 2007, May 2008, Apr 2009, Feb 2010, Jun 2011, Sept 2012, Jun 2013	2006	71	3 MB
				2007	84	5 MB
				2008	143	9 MB
				2009	137	7 MB
				2010	143	7 MB
				2011	190	11 MB
				2012	170	8 MB
				2013	152	8 MB
BD-	http://www.	NaN	NaN	2006	10	248 KB

Diabetes Education	bddiabetes.com					
Centers for Disease Control and Prevention	http://www.cdc.gov/	NaN	NaN	2007	404	16 MB
International Diabetes Federation	http://www.idf.org/	NaN	NaN	2006	20	1 MB

Table 22: The generated queries for “Nephropathy Screening and Treatment” section

Terms	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
diabetes care	3.80	2.19	2.19	2.19	2.19	2.19
type diabetes	3.80	3.80	3.80	3.80		3.80
clinical	2.45	2.45				
treatment	2.45		2.45	2.45	2.45	
assessment	2.19					
complications	2.19	2.19				
etiology	2.19					
exercise	2.19		2.19		2.19	
glucose control	2.19	2.19	2.19	2.19		
kidney	2.19	2.19		3.80	2.19	2.19
damage		2.19		2.19		
kidney disease		2.19			2.19	2.19
metabolic		2.19				
specialist		2.19				
blood glucose		0.84	0.84			3.80
clinical trial		0.84		2.45		
diagnosis			2.19			
monitoring			2.19			2.19
screening			2.19			2.19
therapy			2.19			3.80
blood				2.45		2.45
diabetic				2.45		2.45
risk				2.45		
calcium				2.19		
chronic kidney				2.19	2.19	2.19
chronic kidney disease				2.19	2.19	2.19
complication				2.19	2.19	2.19
diet				2.19	2.19	
kidney disease				2.19		
kidney failure				2.19		
albumin					2.19	2.19
diabetes hypertension					2.19	
diabetes management					2.19	
dialysis					2.19	
esrd					2.19	
hypertension					2.19	
medication					2.19	2.19
urine					2.19	2.19
urine albumin					2.19	2.19

urine albumin excretion						2.19

In the testing phase, generated queries are run over the testing data set given in Table 23 and timeliness prediction is done according to years.

Table 23: The testing data set

Website	**URL**	**HON Monitoring Dates**	**Capture Dates of Archive-it**	**Website Features**		
				Year	**Number of pages**	**Size**
National Diabetes Information Clearinghouse (NDIC)	http://www.diabetes.niddk.nih.gov/	Sep 17 2013, Jan 18 2011, Dec 28 2010, Dec 01 2009, Nov 21 2008, Sep 23 2008, Apr 07 2008, Feb 03 2006	Jun 2008, Jun 2009, Dec 2010, Jun 2011, Jun 2012, Jun 2013	2008	1438	55 MB
				2009	1626	123 MB
				2010	1516	110 MB
				2011	1666	130 MB
				2012	2770	147 MB
				2013	852	37 MB
dLife - For your diabetes life	http://www.dlife.com/	May 18 2012, Dec 08 2011, Nov 01 2011, Sep 15 2011, Sep 12 2011, Aug 26 2011, Aug 21 2011, Sep 24 2010, Sep 21 2010, Apr 30 2010, Mar 29 2010, Mar 22 2010, Mar 31 2009, Mar 20 2009, Dec 01 2008, Oct 07 2008, Sep 24 2008, Apr 23 2008, Mar 27 2008, Nov 14 2007, Jun 19 2006	Apr 2007, Apr 2009, Feb 2010, Jul 2012, Jul 2013	2007	9	448 KB
				2009	52	3 MB
				2010	52	3 MB
				2012	131	10 MB
				2013	201	20 MB
Joslin Diabetes Center	http://www.joslin.org/	Apr 21 2011, Sep 29 2008	Sept 2009, Jul 2011, Aug 2012, Feb 2013	2009	77	4 MB
				2011	125	6 MB
				2012	169	11 MB
				2013	10	5 MB
Diabetes Australia	http://www.diabetesaustralia.com.au	Jun 18 2012, Aug 08 2011, Nov 01 2011, Sep 15 2011, Sep 12 2011, Aug 26 2011, Aug 21 2011, Sep 24 2010,	Jun 2006, Jun 2007, Jun 2008, Jun 2009, Jun 2010, Jun 2011, Jun 2012, Jun 2013	2006	20	92 KB
				2007	35	180 KB
				2008	31	176 KB
				2009	57	316 KB
				2010	58	316 KB
				2011	70	392 KB
				2012	86	2 MB
				2013	187	8 MB

		Sep 21 2010, Apr 30 2010, Mar 29 2010, Mar 22 2010, Mar 31 2009, Mar 20 2009, Dec 01 2008, Oct 07 2008, Sep 24 2008, Apr 23 2008, Mar 27 2008, Nov 10 2006, May 19 2006				
DiabetesMine.com	http://www.diabetesmine.com/	Apr 17 2012, Apr 13 2012, Jun 21 2010, Jun 18 2010, Jun 06 2009, May 31 2009, May 14 2009, May 12 2009, Apr 23 2009, Apr 16 2009, Mar 26 2009	Dec 2010, Jun 2011,Jul 2013	2010	109	14 MB
				2011	221	27 MB
				2013	175	17 MB
Six Until Me	http://www.sixuntilme.com	Feb 01 2012, Jan 20 2012, Dec 28 2011, Dec 06 2011, Nov 08 2011, Sep 26 2011, Aug 03 2011, Jul 13 2011, Apr 11 2011, Apr 06 2011, Jan 20 2011, Jan 05 2011, Nov 30 2010, Oct 13 2009, Nov 04 2008, Oct 30 2008, Oct 27 2008, Sep 17 2008, Sep 03 2008, Aug 30 2008	Nov 2010, Apr 2011, Aug 2013	2010	542	39 MB
				2011	783	77 MB
				2013	171	7 MB

For each section in *RG*, the queries generated for 6 years are run and their scores are obtained.

Table 24: The accuracy of the proposed framework (according to years)

Website	**Number of Captures Crawled**	**Number of Misclassified Captures**	**Accuracy**

http://diabetesaustralia.com.au	8	3	62.5%
http://www.diabetesmine.com/	3	0	100%
http://www.dlife.com/	5	2	60%
http://www.joslin.org/	4	2	50%
http://www.diabetes.niddk.nih.gov/	6	2	66%
TOTAL	**26**	**9**	**65%**

The results are given in Table 24. On average, the proposed methodology succeeds to classify 65% of the websites into the appropriate years. When the misclassified captures are inspected, it is observed that the proposed framework tends to underestimate the year of the website. This could be due to the fact that even though the content of the Web sites is being checked frequently, the contents of the websites are not up to date with the latest research as stated in (Post & Mainous, The Accuracy of Nutrition Information on the Internet for Type 2 Diabetes, 2010). The most common failure is labeling the 2013 websites as 2012 and 2011 websites as 2010 all of which appear in the same clusters in Table 18.

5.7 Timeliness Prediction According to Year Intervals

Post and Mainous argue in their studies that update times displayed at the websites do not guarantee the accuracy of the information and the recent findings are covered in the websites with a significant delay(Post & Mainous, The Accuracy of Nutrition Information on the Internet for Type 2 Diabetes, 2010). In addition, they may be insignificant changes between the consequent years as can be seen in Table 18. Consequently, predictions according to predetermined intervals are also undertaken in this study. The intervals are determined according to Table 18. The cluster indices corresponding to the predicted years are compared with the cluster indices corresponding to the actual years of the web sites. As can be seen from Table 25 the accuracy is increased to 77%.

Table 25: The accuracy of the proposed framework (according to year intervals)

Website	**Number of Captures Crawled**	**Number of Misclassified Captures**	**Accuracy**
http://diabetesaustralia.com.au	8	2	75%
http://www.diabetesmine.com/	3	0	100%
http://www.dlife.com/	5	2	60%
http://www.joslin.org/	4	1	75%
http://www.diabetes.niddk.nih.gov/	6	1	83%
TOTAL	**26**	**6**	**77%**

When the misclassified captures are inspected, it is observed that the proposed framework still tends to underestimate. The websites belonging to cluster 3 are predicted as to belong to cluster 2. This could be due to the fact that we consider many topics regarding diabetes in prediction. As a result, if the authors of a website updates the content of a website for a specific topic only and ignore the others, it causes the framework to underestimate the update time even according to intervals.

5.8 Results and Discussion

The massive growth of health information on the Internet has made the need for tools to query relevant and high quality information vital.

In health domain, although timeliness is reported to be one of the most important dimensions of information quality, current query systems disregard the assessment of it. In this thesis, automatic term recognition, relevance feedback and information retrieval techniques have been used to evaluate content timeliness of diabetes websites according to the current evidence based medicine recommendations. The approach is promising to be replicable and generalizable to other domains. The findings can be utilized by both information seekers and website owners.

CHAPTER 6

DISCUSSION

The previous chapters described the proposed models to build a framework that focuses on information quality and timeliness. This chapter explains the scope of this thesis on the topic of diabetes identifies the limits and discusses the potential application to other domains.

1.1 Diabetes as a central topic

Diabetes is a common health condition from which at least 171 million people suffer in the world. The proposed method might be of benefit to a wider community including web masters, doctors, and other information seekers in diabetes domain. Since there is no publicly available data set to test the proposed methods, many manual processes were involved in the collection and preparation of the data sets in the thesis.

We also considered all the sections of ADA guideline (American Diabetes Association, 2011) in order to generate the queries. The guideline has several sections regarding different aspects of diabetes. To conduct a more sensitive study, it is possible to focus on one of the sections of the guideline which is updated frequently and score the contents based on each subtopic.

1.2 Evidence-based Medicine

There are two main purposes to use EBM in this thesis; in order to represent a gold standard in quality for evaluating the experiment results and to derive quality and timeliness queries for experiments. For the topic of diabetes, the EBM data was obtained from ADA website (American Diabetes Association, 2011).

For other health topics, EBM data can be obtained from several trusted sources such as EBM Online (Evidecen Based Medicine, 2014). Another reliable EBM resource is Database of Abstracts of Reviews of Effectiveness (DARE) (Database of Abstracts of Reviews of Effects (DARE), 2014). DARE contains details of systematic reviews that evaluate the effects of healthcare interventions. Reviews are assessed in means of quality which makes DARE a key resource for busy decision-makers in both healthcare policy and practice.

1.3 Generalization to other domains

The techniques proposed in this thesis are expected to be generalizable across many topics. For many health topics, EBM provides a very specific definition of quality. Consequently, it is promising that health websites regarding other health topics can also be assessed using similar techniques.

The method is scalable to other health topics by using periodically published EBM guidelines for the relevant topic. These reference guidelines are expected to include specialized terms varying in different time intervals:

- To generate timeliness queries, the use of an EBM guideline is essential. Such guidelines are available for some health topics such as depression, arthritis, breast cancer or cardiovascular disease at Cochrane Library (The Cochrane Library , 2014). The Cochrane Collaboration provides reviews for several health conditions based on a comprehensive and expert analysis of the available literature. For those that are not addressed by Cochrane library, EBM Online (Evidecen Based Medicine, 2014) or Database of Abstracts of Reviews of Effectiveness (DARE) (Database of Abstracts of Reviews of Effects (DARE), 2014) can be used to construct "timeliness" queries. These sources contain details of systematic reviews that evaluate the effects of healthcare interventions.
- The extraction of specialized terms plays an important role in evaluating timeliness. If there is no significant update or terminology shift in the referenced guidelines, the method may not able to work efficiently. However, many health topics advance day by day and new terminology (such as new treatments or drugs) emerge and they are widely used for a period of time.

The precision of the query generation phase strongly depends on the term distribution on training websites. If the websites included in the training set have a delay in publishing the latest research, the generated timeliness queries will fail to meet the expectations and to discriminate the up-to-date contents from the out-dated ones. However, the success of any classifier models in machine learning domain relies heavily on training data.

For the general purpose websites which cover several health topics, an approach can be considered by applying the techniques proposed in this thesis for each health topic individually and generating an overall result. However, it may be costly to obtain both evidence based medicine guidelines and low and high quality websites and their archives to build the framework.

In non-health areas, the quality could be harder to define. There may be no clear notion of evidence-based quality. For such areas, it will not be possible to generalize the framework. In addition, this method is scalable to specialized web sites such as diabetes web sites. For general purpose health web sites which cover several health topics, the scalability will be limited.

CHAPTER 7

CONCLUSION AND FUTURE WORK

With the exponential growth of health related information presented on the web, the need for tools to query up-to-date, relevant and high quality information has become paramount. The proposed methodology estimates the information quality and timeliness of diabetes web sites according to EBM with an average accuracy of 76% and 77% respectively. The method is first of its kind in the literature to the best of our knowledge which considers bias content detection, information quality assessment and timeliness assessment at the same time. The contributions of the thesis to the research in the area of Web health information can be listed as follows;

1) An automated quality assessment method
   There are information quality rating tools in the literature which are manual or unreliable and complex to use. Most of them are highly time-consuming or even require expert knowledge which makes them inapplicable for ordinary information seekers. The proposed model is fully automated. However, a further effort is required to make it to be consumer-friendly form in the future.

2) An automated quality assessment method with the functionality of bias content detection
   Previous approaches for information quality assessment in the literature are either manual or are not able to identify bias information on health web sites. Since bias contents may appear on many websites, it is important to incorporate bias content detection functionality in a quality assessment framework.

3) Proposal of a customized dictionary for bias content detection for health related websites
   Recent dictionaries used for opinion mining aid identification of subjectivity but not the bias. Consequently, both the positivity and negativity of the commerce related synsets are zero which makes them inefficient to be used for biased content detection in health domain. Selecting the terms using the relevance feedback algorithm on biased websites and assigning the negativity scores of these synsets a value of 1 dramatically enhanced the precision of bias content detection in the experiments.

4) A model which assesses timeliness of a given web site
   Conventional indicators such as last update time or presence of broken links have shown to be inadequate to express timeliness according to EBM since these indicators do not consider a reference guideline to assess the time interval the content belongs to. Utilization of the terms automatically extracted from periodically published reference guidelines such as ADA as timeliness indicators is the first of its kind in the literature. On the other hand, since it is an automated process, as the new terminology appears, it can easily be adapted to the system without any manual assessment.

The proposed methodology also exceeds the current temporal information retrieval techniques in health domain since the dates given in the text explicitly or implicitly does not guarantee content to reflect up-to-date information. Detection of them as done in conventional temporal information retrieval techniques will be insufficient in health domain. Extraction of the terms given in the guidelines belonging to different time intervals and utilizing them in temporal assessment is a new approach in temporal information retrieval domain.

Despite all these contributions, there are several issues that merit further investigation. For general purpose websites, the proposed model can be applied for each health topic separately and all the timeliness predictions and quality assessments can be combined to produce an overall score. However, it may be costly to obtain both EBM guidelines, low and high quality websites and their archives to build the framework. The method can be improved so as to work with generic web sites.

Another open issue that needs further analysis relates to the weights assigned to the sections while scoring the websites. The method can give weight to each section in the reference guideline differently so as to be in proportion with its importance (how much that section has changed compared to the previous years). By this way, the sections which have changed significantly compared to the other sections in consequent years will have a higher impact in scoring of a web site whereas sections which have not changed at all will have small or no impact.

Although, the study proposed in this thesis focuses on websites, dissemination low quality health information is also possible with social media (Scanfeld, Scanfeld, & Larson, 2010). Rather than high volume websites, the proposed techniques can be applied on social media status updates to detect misuse of drugs or dissemination of invalid information.

In this thesis, Glossex method has been preferred for term recognition as it is simple and fast approach to use. Although successful results are obtained using Glossex, it does not always guarantee to extract terms that are solely related to medicine. As a consequence, the use of existing tools to extract medical concepts such as metaMap (Aronson, 2001) or cTakes (Savova, et al., 2010) can be considered in the future which take into account semantics in the text. An alternative semi-automatic approach can be developed in the future to achieve better and more trustworthy results where the automated part provides filtered initial results to the users.

REFERENCES

[1] Ahmad, K., Gillam, L., & Tostevin, L. (1999). University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). *TREC.*

[2] Alonso, O., Gertz, M., & Baeza-Yates, R. (2009). Clustering and exploring search results using timeline constructions. *18th ACM Conference on Information and Knowledge Management.* ACM.

[3] Alonso, O., Gertz, M., & Baeza-Yates, R. (2009). Clustering and exploring search results using timeline constructions. *18th ACM Conference on Information and Knowledge Management.* ACM.

[4] Alonso, O., Strotgen, J., Baeza-Yates, R. A., & Gertz, M. (2011). Temporal Information Retrieval: Challenges and Opportunities. *In: 1st Temporal Web Analytics Workshop at WWW.*

[5] *American Diabetes Association*. (2011). Retrieved 06 05, 2011, from American Diabetes Association: http://www.diabetes.org/

[6] *American Diabetes Association*. (2011). Retrieved 06 05, 2011, from American Diabetes Assosiation: http://www.diabetes.org/

[7] *American National Corpus*. (2009). (American National Corpus) Retrieved 8 15, 2013, from http://americannationalcorpus.org/OANC/index.html

[8] *Archive-It: A service of the Internet Archive*. (2013). Retrieved 09 01, 2013, from https://archive-it.org/

[9] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium.*

[10] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium* (p. 17). American Medical Informatics Association.

[11] Bansal, M., Cardie, C., & Lee, L. (2008). The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. *Proceedings of COLING: Companion volume: Posters*, 1-16.

[12] Becchetti, L., Castillo, C., Donato, D., Leonardi, S., & Baeza-Yates, R. (2006). Using rank propagation and probabilistic counting for link-based spam detection. *Proc. of WebKDD, 6.*

[13] Belen Sağlam, R., & Taşkaya Temizel, T. (2014, Jan 6). A framework for automatic information quality ranking of diabetes websites. *Informatics for Health and Social Care*.

[14] Bernstam, E., Shelton, D., Walji, M., & Meric-Bernstam, F. (2005). Instruments to assess the quality of health information on the World Wide Web: what can our patients actually use? *International journal of medical informatics, 74*(1), 13-19.

[15] Bouzeghoub, M. (2004). A framework for analysis of data freshness. In *Proceedings of the 2004 international workshop on Information quality in information systems* (pp. 59-67). ACM.

[16] Bouzeghoub, M. (2004). A framework for analysis of data freshness. In *Proceedings of the 2004 International Workshop on Information quality in Information Systems* (pp. 59-67). ACM.

[17] Clark, E. J. (2002). Health Care Web Sites: Are They Reliable? *Journal of Medical Systems, 26*(6), 519-528.

[18] Das, S., & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. *In Asia Pacific Finance Association Annual Conf. (APFA).*

[19] *Data Mining Tools Used Poll*. (2013). (KDnuggets) Retrieved 10 15, 2013, from http://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm

[20] *Database of Abstracts of Reviews of Effects (DARE)*. (2014). (The Cochrane Collaboration) Retrieved May 10, 2014, from http://www.cochrane.org/editorial-and-publishing-policy-resource/database-abstracts-reviews-effects-dare

[21] Dave, K., Lawrence, S., & Pennock, D. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web* (pp. 519--528). ACM.

[22] Denecke, K. (2008). Accessing medical experiences and information. *European Conference on Artificial Intelligence, Workshop on Mining Social Data.*

[23] Denecke, K. (2009). Are SentiWordNet scores suited for multi-domain sentiment classification? *Fourth International Conference on Digital Information Management (ICDIM 2009)* (pp. 1--6). IEEE.

[24] Denecke, K., & Bernauer, J. (2007). Extracting specific medical data using semantic structures. *Artificial Intelligence in Medicine, 11th Conference on Artificial Intelligence in Medicine, AIME 2007* (pp. 257-264). Amsterdam, The Netherlands: Springer, 978-3-540-73598-4.

[25] *DMOZ open directory project*. (2014). Retrieved 03 18, 2014, from http://www.dmoz.org/

[26] Dutta-Bergman, M. (2004). Health attitudes, health cognitions, and health behaviors among Internet health information seekers: Population-based survey. *Journal of Medical Internet Research, 6*(2).

[27] *Evidecen Based Medicine*. (2014). (BMJ Publishing Group Ltd.) Retrieved May 10, 2014, from http://ebm.bmj.com/

[28] Eysenbach, G., & Kohler, 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. *British Medical Journal*, 573.

[29] Fallis, D., & Frick, M. (2002). Indicators of accuracy of consumer health information on the Internet. *Journal of the American Medical Informatics Association*, 73.

[30] Fox , S., & Rainie, L. (2002, 5 22). *Vital Decisions: A Pew Internet Health Report.* Retrieved from PewResearch Internet Project: http://www.pewinternet.org/2002/05/22/vital-decisions-a-pew-internet-health-report/

[31] *General Inquirer Home Page*. (2002, 5 5). Retrieved from General Inquirer: http://www.wjh.harvard.edu/~inquirer/

[32] Geng, G.-G., Zhang, X.-C., Jin, X.-B., & Zhang, D.-X. (2010). Evaluating web content quality via multi-scale features. *Proceedings of the ECML/PKDD.* ECML/PKDD.

[33] Gertz, M., Özsu, M., Saake, G., & Sattler, K.-U. (2004). Report on the dagstuhl seminar. *ACM SIGMOD Record, 33*(1), 127-132.

[34] Goldberg, A., Zhu, X., & Wright, S. (2007). Dissimilarity in graph-based semi-supervised classification. *Artificial Intelligence and Statistics (AISTATS).*

[35] Griffiths, K., Tang, T. T., Hawking, D., & Christensen, H. (2005). Automated assessment of the quality of depression websites. *Journal of medical Internet research, 7*(5).

[36] Griffiths, K., Tang, T. T., Hawking, D., & Christensen, H. (2005). Automated assessment of the quality of depression websites. *Journal of Medical Internet Research, 7*(5).

[37] Griiffiths, K., Tang, T. T., Hawking, D., & Christensen, H. (2005). Automated assessment of the quality of depression websites. *Journal of medical Internet research, 7*(5).

[38] Gyöngyi, Z., Garcia-Molina, H., & Pedersen, J. (2004). Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30* (pp. 576-587). VLDB Endowment.

[39] Health on the Net Foundation. (2013, 6). *Health on the Net Foundation*. Retrieved August 23, 2010, from Health on the Net Foundation: http://www.hon.ch/index.html

[40] Hibbard, J. (2003). Engaging health care consumers to improve the quality of care. *Medical care*.

[41] HON. (2010, February 8). *Health On the Net Foundation*. Retrieved August 23, 2010, from Health on the Net Foundation: http://www.hon.ch/index.html

[42] HON. (2010, February 8). *Health On the Net Foundation*. Retrieved August 23, 2010, from Health on the Net Foundation: http://www.hon.ch/index.html

[43] Hopkins, D., & King, G. (2007). Extracting systematic social science meaning from text. *Manuscript available at http://gking.harvard.edu/files/words.pdf*.

[44] *Internet Archive*. (2001, March 10). Retrieved 11 10, 2013, from https://archive.org/

[45] Jin, X., Li, Y., Mah, T., & Tong, J. (2007). Sensitive webpage classification for content advertising. *Proceedings of the International Workshop on Data Mining and Audience Intelligence for Advertising.*

[46] Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation, 28*(1), 11--21.

[47] Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology, 3*(2), 259-289.

[48] Kanhabua, N., & Norvag, K. (2009). Using temporal language models for document dating. In *Machine Learning and Knowledge Discovery in Databases* (pp. 738--741). Springer.

[49] Kanhabua, N., & Norvag, K. (2009). Using temporal language models for document dating. In *Machine Learning and Knowledge Discovery in Databases* (pp. 738--741). Springer.

[50] Keikha, M., Gerani, S., & Crestani, F. (2011). Temper: A temporal relevance feedback method. In *Advances in Information Retrieval* (pp. 436--447). Springer.

[51] Kim, P., Eng, T. R., Deering, M., & Maxfield, A. (1999). Published criteria for evaluating health related web sites: review. *British Medical Journal, 318*(7184).

[52] Klimov, D., Shahar, Y., & Taieb-Maimon, M. (2010). Intelligent visualization and exploration of time-oriented data of multiple patients. *Artificial intelligence in medicine, 49*(1), 11--31.

[53] Klimov, D., Shahar, Y., & Taieb-Maimon, M. (2010). Intelligent visualization and exploration of time-oriented data of multiple patients. *Artificial Intelligence in Medicine, 49*(1), 11--31.

[54] Knoth, P., Schmidt, M., Smrz, P., & Zdrahal, Z. (2009). Towards a framework for comparing automatic term recognition methods.

[55] Knoth, P., Schmidt, M., Smrz, P., & Zdrahal, Z. (2009). Towards a framework for comparing automatic term recognition methods. *Conference Znalosti 2009.* Brno, Czech Republic. .

[56] Korkontzelos, I., Klapaftis, I., & Manandhar, S. (2008). Reviewing and evaluating automatic term recognition techniques. In *Advances in Natural Language Processing* (pp. 248-259). Springer.

[57] Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, Y., & Cofino, T. (2004). Glossary extraction and utilization in the information search and delivery system for IBM Technical Support. *IBM Systems Journal, 43*(3), 546--563.

[58] Lambert, V. (2010, 10 15). *Finding health information on the internet*. Retrieved from The Telegraph: http://www.telegraph.co.uk/health/wellbeing/8066878/Finding-health-information-on-the-internet.html

[59] Lin, Y.-K., Chen, H., & Brown, R. (2013). MedTime: A temporal information extraction system for clinical narratives. *Journal of biomedical informatics, 46*, S20--S28.

[60] Lin, Y.-K., Chen, H., & Brown, R. A. (2013). MedTime: A temporal information extraction system for clinical narratives. *Journal of Biomedical Informatics, 46*, S20-S28.

[61] *looksmart*. (2014). Retrieved 03 18, 2014, from http://www.looksmart.com/

[62] Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

[63] Marshall, L., & Williams, D. (2006). Health information: does quality count for the consumer? *Journal of Librarianship and Information Science, 38*(3), 141.

[64] *MedlinePlus Guide to Healthy Web Surfing*. (2010, 08 11). Retrieved 06 06, 2011, from MedlinePlus: http://www.nlm.nih.gov/medlineplus/healthywebsurfing.html

[65] Miyanishi, T., & Sakai, T. (2013). Time-aware structured query suggestion. *36th international ACM SIGIR Conference on Research and Development in Information Retrieval.*

[66] Miyanishi, T., & Sakai, T. (2013). Time-aware structured query suggestion. *36th international ACM SIGIR Conference on Research and Development in Information Retrieval.*

[67] Moreno, J., Del Castillo, J., Porcel, C., & Herrera-Viedma, E. (2010). A quality evaluation methodology for health-related websites based on a 2-tuple fuzzy linguistic approach. *Soft Computing-A Fusion of Foundations, Methodologies and Applications, 14*(8), 887-897.

[68] Morr, S., Shanti, N., Carrer, A., Kubeck, J., & Gerling, M. (2010). Quality of information concerning cervical disc herniation on the Internet. *The Spine Journal, 10*(4), 350-354.

[69] Moyer, C. S. (2012, 1 30). *Cyberchondria: the one diagnosis patients miss*. (American Medical News) Retrieved 10 30, 2012, from American Medical News: http://www.ama-assn.org/amednews/2012/01/30/hll10130.htm

[70] National Institutes of Health's Web Site. (2012, 4). *MedlinePlus Guide to Healthy Web Surfing*. Retrieved 06 06, 2011, from MedlinePlus: http://www.nlm.nih.gov/medlineplus/healthywebsurfing.html

[71] Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR), 33*(1), 31-88.

[72] Norman, F. (1998). Organizing medical networked information (OMNI). *Informatics for Health and Social Care, 23*(1), 43--51.

[73] Ntoulas, A., Najork, B., Manasse, M., & Fetterly, D. (2006). Detecting spam web pages through content analysis. *roceedings of the 15th international conference on World Wide Web* (pp. 83--92). ACM.

[74] Page, L., & Brin, S. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the 7th international conference on World Wide Web (WWW). 30*, pp. 107-117. Brisbane: Australia.

[75] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*, 1-135.

[76] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 79-86).

[77] Peetz, M.-H., Meij, E., de Rijke, M., & Weerkamp, W. (2012). Adaptive temporal query modeling. In *Advances in Information Retrieval* (pp. 455--458). Springer.

[78] Peralta, V. (2006). Data freshness and data accuracy: A state of the art. *Instituto de Computacion, Facultad de Ingenieria, Universidad de la Republica, Uruguay, Tech. Rep. TR0613*.

[79] Porter, M. (2006, 1). *The Porter Stemming Algorithm*. Retrieved 2 14, 2013, from Porter Stemmer: http://tartarus.org/martin/PorterStemmer/

[80] Post, R., & Mainous III, A. (2010). The Accuracy of Nutrition Information on the Internet for Type 2 Diabetes. *Archives of Internal Medicine, 170*(16), 1504.

[81] Post, R., & Mainous, A. G. (2010). The Accuracy of Nutrition Information on the Internet for Type 2 Diabetes. *Archives of internal medicine, 170*(16), 1504.

[82] *RapidMiner*. (n.d.). Retrieved 12 03, 2011, from RapidMiner: http://rapid-i.com/content/view/181/190/lang,en/

[83] *RapidMiner*. (2014). Retrieved 12 03, 2011, from RapidMiner: http://rapid-i.com/content/view/181/190/lang,en/

[84] Resnick, P., & Miller, J. (1996). PICS: Internet access controls without censorship. *Communications of the ACM, 39*(10), 87-93.

[85] Riloff, E., Wiebe, J., & Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. *Proceedings of AAAI*, (pp. 1106–1111).

[86] Robertson, S. (1990). Documentation Note on Term Selection for Query Expansion. *Journal of documentation, 46*, 359--364.

[87] Robertson, S., & Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science, 27*(3), 129--146.

[88] Robertson, S., & Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129-146.

[89] Robertson, S., & Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 232-241). Springer-Verlag New York, Inc.

[90] Robertson, S., & Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 232-241). Springer-Verlag New York, Inc.

[91] Rocchio, J. (1971). Relevance feedback in information retrieval.

[92] Sackett, D., Rosenberg, W., Gray, J., Haynes, R., & Richardson, W. (1996). Evidence based medicine: what it is and what it isn't. *BMJ: British Medical Journal, 312*(7023), 71.

[93] Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613--620.

[94] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association, 17*(5), 507-513.

[95] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 507--513.

[96] Scanfeld, D., Scanfeld, V., & Larson, E. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control*, 182--188.

[97] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the international conference on new methods in language processing*, (pp. 44--49). Manchester.

[98] Schmid, H. (n.d.). *TreeTagger*. Retrieved 02 02, 2012, from http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[99] Scullard, P., Peacock, C., & Davies, P. (2010). Googling children's health: reliability of medical advice on the internet. *Archives of Disease in Childhood, 95*(8), 580-582.

[100] Seidman, J., Steinwachs, D., & Rubin, H. (2003). Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites. *Journal of Medical Internet Research, 5*(4), e30.

[101] Seidman, J., Steinwachs, D., & Rubin, H. (2003). Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites. *Journal of Medical Internet Research, 5*(4), e30.

[102] *SentiWords*. (2013). Retrieved 5 5, 2014, from Human Language Technology: https://hlt.fbk.eu/technologies/sentiwords

[103] Shankaranarayan, G., Ziad, M., & Wang, R. (2003). Managing data quality in dynamic decision environments: An information product approach. *Journal of Database Management (JDM), 14*(4), 14-32.

[104] Shin, B. (2003). An Exploratory Investigation of System Success Factors in Data Warehousing. *Journal of the Association for Information Systems, 4*(1).

[105] Sillence, E. a. (2006). A framework for understanding trust factors in web-based health advice. *International journal of human-computer studies, 64*(8), 697--713.

[106] Stvilia, B., Mon, L., & Yi, Y. (2009). A model for online consumer health information quality}. *Journal of the American Society for Information Science and Technology, 60*(9), 1781--1791.

[107] *Terrier IR Platform v3.5*. (2014). Retrieved 2 14, 2013, from University of Glasgow School of Computing Science: http://terrier.org/

[108] *The Cochrane Library* . (2014). Retrieved May 25, 2014, from http://www.thecochranelibrary.com/view/0/index.html

[109] *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)* . (2010, May 03). Retrieved 06 04, 2011, from ECML PKDD 2010: http://www.ecmlpkdd2010.org/

[110] Theodoratos, D., & Bouzeghoub, M. (1999). Data Currency Quality Factors in Data Warehouse Design. In *DMDW* (p. 15).

[111] Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 327-35). Association for Computational Linguistics.

[112] Thorpe, B., Kiebzak, G., Chavez, J., Lewiecki, E., & Rudolph, L. (2006). Assessment of osteoporosis-website quality. *Osteoporosis international, 17*(5), 741-752.

[113] Tong, R. M. (2001). An operational system for detecting and tracking opinions in on-line discussion. *Proceedings of the Workshop on Operational Text Classification (OTC),*.

[114] Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification. *Proceedings of the Association for Computational Linguistics (ACL)*, (pp. 417-424).

[115] University of Glasgow School of Computing Science. (n.d.). *Terrier IR Platform v3.5*. Retrieved 2 14, 2013, from http://terrier.org/

[116] University of Oxford, Division of Public Health and Primary Health Care. (2010). *DISCERN-Quality criteria for consumer health information*. Retrieved 11 20, 2010, from DISCERN: http://www.discern.org.uk

[117] van der Marel, S., Duijvestein, M., Hardwick, J., van den Brink, G., Veenendaal, R., Hommes, D., & Fidder, H. (2009). Quality of web-based information on inflammatory bowel diseases. *Inflammatory bowel diseases, 15*(12), 1891-1896.

[118] Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *J. of Management Information Systems, 12*(4), 5-33.

[119] Wang, Y., & Liu, Z. (2007). Automatic detecting indicators for quality of health information on the Web. *International Journal of Medical Informatics, 76*(8), 575-582.

[120] *What is LIWC?* (2014, 5 5). Retrieved from Linguistic Inquiry and Word Count: http://www.liwc.net/

[121] Willett, P. (2006). The Porter stemming algorithm: then and now. *Program: electronic library and information systems, 40*(3), 219--223.

[122] Willett, P. (2006). The Porter stemming algorithm: then and now. *Program: Electronic library and Information Systems*, *40*(3), 219-223.

[123] Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. *Proceedings of the IEEE International Conference on Data Mining (ICDM).*

[124] Yuanhua, L., & ChengXiang, Z. (2011). Lower-bounding term frequency normalization. *Proceedings of CIKM'2011*, (pp. 7-16).

[125] Zhang, E., & Zhang, Y. (2006). UCSC on TREC 2006 blog opinion mining. *Proceedings of the Text Retrieval Conference (TREC).*

[126] Zhang, Z., Iria, J., Brewster, C., & Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC2008).* Marrakesh, Morocco.

[127] Zhang, Z., Iria, J., Brewster, C., & Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC2008).* Marrakesh, Morocco.

APPENDIX A

EVOLUTION of ADA GUIDELINES

Table 26: The number of words in ADA guidelines

	2006	2007	2008	2009	2010	2011	2012	2013
Assessment Of Common Comorbid Conditions							969	995
Bariatric Surgery				550		641	718	724
Classification And Diagnosis	532	917	956	731	2755	1796	1973	2038
CVD	3957	3957	4037	4576	5384	6014	5623	6139
Detection And Diagnosis of GDM	406	490	678	683	884	850	997	1169
Diabetes And Employment				4317	4595	4595	4572	4572
Diabetes Care In Specific Populations	3104	3305	3685	3794	4045	4437	4761	4911
Diabetes Care In Specific Settings	4723	5483	5864	6368	4091	3714	4042	4150
Diabetes And Driving							6523	6531
Diabetes Care In The School And Day Care	3224	4037	4069	4416	4471	4476	4449	4445
Diabetes Management In Correctional İnstitutions	6144	6175	6018	6050	6057	6059	6060	6091
Diagnosis And Classification Of Diabetes	4670	4671	4663	4655	7471	7136	7096	7294
DSME	364	366	408	463	623	651	665	874
Foot Care	818	897	978	1124	1110	1011	1016	1018
Glycemic Control	1753	2106	3069	5824	6078	4024	4319	4496
Hypoglycemia And Employment/Licensure	610	612	612					
Hypoglycemia	297	296	501	480	472	566	587	894
İmmunization	300	307	299	298	298	274	329	642
Initial Evaluation	1266	561	876	279	281	866	280	590
Intercurrent İllness	211	211	185	178	184	184	184	176
MNT	2531	767	1461	1602	2350	1565	1420	1491
Nephropathy Screening And Treatment	1826	1735	1414	1583	1634	1732	1604	1677

Neuropathy Screening And Treatment	1089	980	882	785	884	825	851	980
Nutrition Recommendations		16963	16861					
Pharmacologic And Overall Approaches To Treatment						836	1042	955
Physical Activity	1208	1198	823	812	976	982	1041	1035
Prevention/Delay Of Type 2 Diabetes	1504	740	749	1033	923	817	535	591
Psychosocial Assessment And Care	458	448	389	401	401	387	433	484
Retinopathy Screening And Treatment	1067	1065	1088	1094	1164	1148	142	1085
Strategies For İmproving Diabetes Care	777	772	776	792	780	1003	989	976
Testing For Prediabetes and Diabetes in Asymptomatic Patients			1645	1392	1254	1467	1567	1798
Third-Party Reimbursement For Diabetes Care, Self-management Education, And Supplies	1484	1486	1486	1516	1515	1515	1518	1567
Third-Party Reimbursement For Diabetes Care, Self-management Education, And Supplies (in Standard of Diabetes Care)	339	346	332	320				
When Treatment Goals Are Not Met			104	112	144	144	157	121

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Belen Sağlam, Rahime

Nationality: Turkish (TC)

Date and Place of Birth: 15 May 1983, Ankara

Marital Status: Married

Phone: +90 533 7047710

email: rahimebelen@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Information Systems	2009
BS	Başkent University Computer Eng.	2006
High School	Adana Science High School, Adana	2001

WORK EXPERIENCE

Year	Place	Enrollment
2006-2010	METU Informatics Institute. /TURKEY	Research Assistant
2010-2011	Siemens E.C. /TURKEY	Software Engineer
2011-2013	Fujitsu Technology Solutions /TURKEY	Software Engineer, Business Intelligence Expert
2014-present	AYESAŞ /TURKEY	Software Development Team Leader

FOREIGN LANGUAGES

Advanced English

PUBLICATIONS

1. Book Chapters

1. R. Belen and T. Taşkaya Temizel (2010). A Framework to Detect Disguised Missing Data. (A.V.Senthil Kumar, Ed.).Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains.

2. Journal Articles

2. R. Belen Sağlam and T. Taşkaya Temizel, "A framework for automatic information quality ranking of diabetes websites," *Informatics for Health and Social Care,* 6 Jan 2014.

3. Conference Papers

1. R. Belen, T. Taşkaya Temizel and Ö. Kaygısız, "A data quality case study for Turkish highway accident data sets," in *Road safety on four continents: 15th international conference*, Abu Dhabi, United Arab Emirates, 2010.