Semantics-based Automated Quality Assessment of Depression Treatment Web Documents

by

Yanjun Zhang
Graduate Program in Library and Information Science

A thesis submitted in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy

School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

The past decade has witnessed a dramatic expansion in the amount of publicly available health care information on the Web. The health care information on the web, however, is of extremely variable quality. The evaluation of content quality is a big challenge because non-automated methods for information content rating can be easily overwhelmed by the huge data volume. This study proposes an automated approach for assessing the quality of web health care information through comparing the text content with evidence-based health care recommendations. This method relies on semantic analysis and text classification to identify the presentation of evidence-based recommendations in web documents. As a result, the semantics-based rating approach is able to rate quality based on information content, rather than using indirect quality indicators such as website authorship, sponsorship, or text keywords as used in previous studies. Two systems were built to implement the semantics-based quality rating: a rule-based system and a prototypical machine learning system. The performance of both implementations was evaluated by comparing the automated quality rating results with human rating results on the same set of depression treatment web pages. The evaluation demonstrates that the automatically generated rating results using the semantics-based approach are comparable to those from human raters: that is, there is a high Pearson correlation between computer ratings and human rating results. The semantics-based approach has an advantage over previous automated approaches in that it produces quality rating results that present to information consumers feedback that is more instructive than just a quality score.

Keywords: Health Care Information Quality on the Web, Information Quality Assessment, Evidence-based Health care information, Automated Quality Rating, Semantic Analysis, Text Classification

Acknowledgements

This project would not have been possible without the great support from my thesis advisory committee. I would like to thank my supervisors Dr. Jacqueline Burkell and Dr. Robert Mercer for guiding me along this research, contributing thoughtful comments, and providing editorial assistance. Their knowledge, critical thinking and research attitude were invaluable to my research. I would also like to particularly thank my other supervisory committee member, Dr. Hong Cui, for her ideas, her enthusiasm, and her guide along the thesis research. She gave me a lot of encouragement to complete my PhD research and she consistently provided me full support from the initial stage when the research problem was identified till the final stage of thesis editing.

Thanks to the Faculty of Information & Media Studies for all forms of support. Thanks to Dr. Liwen Vaughan for teaching me to conduct independent research and mentoring me during my course study. Thanks to Dr. Isola Ajiferuke for keeping his office always open to me for discussions. Thanks to Dr. Victoria Rubin for guiding me to enter the Natural Language Processing field.

Thanks to the Social Sciences and Humanities Research Council (SHHRC) of Canada for offering me the doctoral fellowships.

Thanks to my parents for their support to my decision to pursue my doctoral studies. A special thank you to my dear wife Xinxin who endured this long process with me and always offered me love, support, and lots of constructive comments on my thesis as well. Without her consistent support and inspiration, this project would not have been completed.

Dedication

To Lucas

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

Abbreviations	Meaning
AHRQ	Agency for Healthcare Research and Quality
AOL	America Online
API	Application Programming Interface
ARFF	Attribute-Relation File Format. ARFF is a type of file format used in WEKA. An ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes.
BHIA	British Healthcare Internet Association
CONANN	A online biomedical concept annotator developed by Lawrence H. Reeve and Hyoil Han in 2007
DARPA	Defense Advanced Research Projects Agency
DISCERN	DISCERN is a brief questionnaire which provides users with a valid and reliable way of assessing the quality of written information on treatment choices for a health problem. The DISCERN project was funded by The British Library and NHS Executive Research & Development Programme
DUC	Document Understanding Conference
EBMWG	Evidence-Based Medicine Working Group
HON	Health On the Net Foundation - a not-for-profit organization founded in 1995 under the auspices of the Geneva Ministry of Health and based in Geneva, Switzerland. One of the first organizations to guide both lay users and medical professionals to reliable sources of health care information in cyberspace.
HONcode	Health On the Net Code of Conduct - a not-for-profit organization founded in 1995 under the auspices of the Geneva Ministry of Health and based in Geneva, Switzerland
HTML	HyperText Markup Language
IHTSDO	International Health Terminology Standards Development Organization
LSA	Latent Semantic Analysis

LVG	Lexical Variant Generator. The Lexical Variant Generation programs perform lexical transformations to text.
MedCERTAIN	MedPICS Certification and Rating of Trustworthy Health care information on the Net
MeSH	Medical Subject Headings
NLM	National Library of Medicine
MMTx	MetaMap Technology Transfer
NHS	National Health Service
NICE	National Institute for Health and Clinical Excellence
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
POS	Part of speech
SKR	Semantic Knowledge Representation
SNOMED CT	Systematized Nomenclature of Medicine--Clinical Terms. It is a systematically organised computer processable collection of medical terms providing codes, terms, synonyms and definitions covering diseases, findings, procedures, microorganisms, substances, etc.
SPECIALIST	The SPECIALIST Natural Language Processing (NLP) Tools. They have been developed by the The Lexical Systems Group of The Lister Hill National Center for Biomedical Communications to investigate the contributions that natural language processing techniques can make to the task of mediating between the language of users and the language of online biomedical information resources. The SPECIALIST NLP Tools facilitate natural language processing by helping application developers with lexical variation and text analysis tasks in the biomedical domain. The NLP Tools are open source resources distributed subject to certain terms and conditions.
TREC	Text Retrieval Conferences
UMLS	Unified Medical Language System. It is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.

URAC	Utilization Review Accreditation Commission - a nonprofit organization promoting health care quality by accrediting health care organizations.
URL	Uniform Resource Locator
WEKA	Weka is a collection of machine learning algorithms for data mining tasks written in Java, developed at the University of Waikato, New Zealand
XML	eXtensible Meta Language

Chapter 1
Introduction

The last decade has witnessed a dramatic expansion in the amount of publicly available health care information on the World Wide Web, and the use of web health care information has become popular among both health care professionals and patients. Recent national surveys in the United States show that 80% of online users look for advice or information about health or health care (Pew Internet and American Life Project, 2011). In fact, this percentage has been consistently above 80% since 2003 (Pew Internet and American Life Project, 2003), even though the Internet user population has grown 33% from 184,447,987 in September 2003 to 245,203,319 in March 2011 as reported by Nielsen//NetRatings Inc. (Internet News, 2003; Internet World Stats, 2012). Moreover, survey results show that younger generations are increasingly turning to the Internet for health care information and it appears that the use of the web to look for health care information will become even more popular among future generations: in the United States, only 73% of senior adults aged more than 65 look for online health care information, but 84% of adults aged from 30 to 49 do so (Pew Internet and American Life Project, 2011).

The World Wide Web has a lot of advantages for knowledge dissemination, including widespread access, interactive interface, and fast content update and circulation. Therefore, it is not surprising that health care information on the web has caused a shift in the ways in which people consume health and medical information, with many patients looking for information online before talking with their physicians (Hesse et al., 2005). Another study (Podichetty et al., 2006) shows that between 23% and 31% of survey respondents from a

cross section of health care professionals reported that more than 80% of patients in their daily practice used the web to find health care information.

Although online health care information is widely used, the quality of health care information on the web is extremely variable in terms of accuracy, coverage and currency (Eysenbach et al. 2002; Kunset et al. 2002; Griffiths & Christensen, 2005). Since the World Wide Web provides an open platform for publishing information, any information provider, including pharmaceutical industry-sponsored organizations, medical experts, and patients, can freely post health care information for different purposes such as advertising, education, or simply the narration of personal experiences. In addition, the maintenance of website health content varies with respect to when it was "last updated" or "last reviewed". Given these factors, the uneven quality of the health care information available on the web is not surprising. Since inaccurate or biased health care information can be widely disseminated through the web to anyone, including caregivers and patients, misinformation on the web could cause and indeed has caused life-threatening accidents (Crocco et al, 2002; Kiley, 2002). According to a recent survey by Pew Internet and American Life Project (2011), for example, although 30% of adults in the U.S. said they or someone they know has been helped by following medical advice or health care information found online, there are also 3% of respondents who indicated that they or someone they know has been harmed by doing so. Because of the potential harm caused by inaccurate information, the quality assessment of health care information on the web has become a common interest of various health care information stakeholders, including e-health policy makers, information providers/consumers, information search service providers, etc.

Despite the great range in the quality of health care information on the web, information consumers themselves make surprisingly little effort to ensure that the information found on the web is of high quality (Eysenbach & Köhler, 2004). Studies indicate that consumers rarely verify information sources or read disclaimers on the websites they use (Pew Internet and American Life Project, 2006). Instead of accessing health care information exclusively from credible web sources, most consumers are accustomed to querying web search engines like Google and just reading the top URLs in the retrieved item list (Peterson et al., 2003; Eysenbach & Köhler, 2004). Moreover, over 50% of consumers access information from unfamiliar websites (Pew Internet and American Life Project, 2006).

Measurements that rate the quality of the web health care information content could assist online consumers to access a higher grade of health care information. The extremely large amount of health care information on the web, however, easily overwhelms the capacity of any manual evaluation system. Therefore, in order to address the large-scale problem of online health care information content quality, we require automated quality assessment instruments that ideally can perform as well as a human rater in evaluating health care information quality.

Much research has been done in related areas, including efforts to develop the definition of high quality online health care information (e.g. Bopp & Smith, 2000), to establish quality rating codes and indicators (e.g. Eysenbach et al., 2002; Griffiths, 2002; URAC, 2007), and to explore rating automation (Griffiths et al., 2005; Hawking et al., 2007; Wang & Liu,

2007). The current study explores a new automated approach, i.e., semantics-based quality assessment, to rate the information quality of web-based health care information documents based on their content. This approach is designed particularly for dealing with text-based online documents. Other types of web document such as multi-media web pages are not in the scope of this study. In contrast to previous studies, the semantics-based approach used in this research accomplishes quality assessment through semantically comparing the web text content with evidence-based health care practice guidelines and rating the quality based on the degree of concordance and the coverage of the best evidence.

The structure of this thesis is as follows. In Chapter 2, previous studies about web health care information quality evaluation are reviewed. Based on this review, the research goals and scope of this study are defined. Chapters 3 and 4 describe the research methodologies. Chapter 3 introduces the experimental design for rating the quality of health care information content, including data sampling, quality rating standards, the process of human rating and automated rating, etc. Chapter 4 gives a detailed description of the design and development of the automated quality assessment system. In Chapter 5, the automated quality rating results are presented and then evaluated by comparing them to human rating results. Finally, Chapter 6 presents the research conclusions based on an analysis of the study results. A discussion of the study limitations is also included in this chapter.

Chapter 2
Literature Review

Research on the quality of online health care information and evaluation started in the late nineties. At that time, Silberg et al. (1997) provided an overview of the quality control and quality assurance issues for medical information on the web. Since then, two basic strategies have been explored to address the issue of web health care information quality: codes of conduct for health care information providers, and the evaluation of health care information content. In practice, the strategies are often combined to improve access to high quality health care information on the web. The first approach specifies basic criteria to which health websites should adhere. These codes of conduct usually include accountability and other criteria (e.g. HONcode, 2012) and provide valuable guidelines for information producers to follow. Codes of conduct, however, are not a silver bullet for solving all the quality issues. First of all, since the Internet is a heterogeneous platform where any information provider can publish information, such types of codes are not enforced on all web sources related to health care information. Without enforcement the warrant is flawed. In addition, a HONcode seal on a certified health web site, for example, attests to an intention to contribute to quality medical information (HONcode, 2012), but it does not speak to the content quality per se. Thus, there is still need for the evaluation of health care information content, often addressed by quality rating schemes.

The challenges for these quality rating schemes are to identify appropriate rating criteria (e.g., subjective vs. objective criteria) and develop effective rating processes (e.g., manual vs. automated processes). One particular issue is the extremely large volume of health care

information on the web that can easily overwhelm non-automated rating efforts. Therefore, it is necessary to explore automated quality rating approaches in order to find a practical solution to the problem of health care information quality rating. The objective of this study is to develop and test an automated process for the evaluation of health care information on the web. This study uses semantic parsing technology to empower computer programs with the capability to parse texts and to apply shallow semantic analysis to sentences in order to rate the quality of health care information based on text semantics.

2.1 A Working Definition for Quality of Health Care Information on the Web

The automated quality assessment instrument developed in this study aims to evaluate the content quality of health care information on the web. Therefore, the automated tool does not rate representational and accessibility qualities such as ease of understanding, aesthetics in web page design, and site navigation. Four quality properties intrinsic to the information content (*content correctness*, *comprehensiveness*, *bias-free* and *content currency*) are considered essential for high quality health care information in many previous Internet health care information quality studies (as referred in the following paragraphs) and are used in well-known Internet health care information quality initiatives such as MedCERTAIN (Eysenbach et al., 2001), URAC (2007), HONcode (2012), etc.

Content correctness is also often referred to as information accuracy (e.g. Frické & Fallis, 2004). The importance of content correctness is self-evident, since health care information consumers use health care information on the World Wide Web for activities such as self-diagnosis and self-treatment, or to determine whether to visit a health care professional. A

survey by Pew Internet and American Life Project (2006) shows that half of online health searches have an impact on people's own health care routine or the way they care for someone else. This principle is also claimed in the quality assessment framework proposed by Eysenbach and Diepgen (1998) as "first, do no harm."

Content comprehensiveness is another key indicator for health care information quality. Take medical treatment information as an example: Charnock et al. (1999) indicated that patients or health care information consumers want to be aware not only of treatment methods, but also of benefits, side-effects, treatment restrictions, and other related information. Obviously, web sources that provide health care information about only one or a few of these aspects may not meet the full information needs of consumers. In a systematic review of empirical studies assessing the quality of health care information on the web (Esynbach et al., 2002), a large number of quality assessment studies focused on evaluating the content completeness of health care information.

Third, health care information content of high quality should be free of subjective bias. Undoubtedly, it is beneficial for consumers to be informed of different treatment options and be provided with clear descriptions of both strengths and drawbacks of each option. Sources demonstrating bias, potentially including those sponsored by the pharmaceutical industry, may advertise some products while not mentioning alternative treatments. In a 2003 study (Bouchier & Bath, 2003) evaluating websites that provide information on Alzheimers' disease, for example, researchers found that a number of websites contained biased information that could mislead readers. Some quality assessment practices try to address the

concern of content bias in their evaluation. For example, the code of conduct defined in HONcode (2012) requires web sources to provide financial disclosure to assist information users to identify potential conflicts of interest and bias.

Content currency is also an important facet of content quality for health care information. High quality web sources should be replenished with up-to-date health literature in a timely manner, and obsolete information should be removed on a regular basis. Unfortunately, it is quite common that information content from some online health care information sites does not reflect or may even conflict with the latest findings in the health care literature, due either to the limited knowledge of information providers or to the lack of content maintenance. Content currency of the health care information is a commonly accepted criterion in web-based health care information evaluation studies (e.g. Bath & Bouchier, 2003; Anderson et al., 2009).

2.2 Quality Indicators and Assessment Approaches

The basic approach to assessing the quality of health care information is to organize a group of medical specialists to carefully review health care information content and rate the quality of that content based on their medical knowledge in specific domains. This method, however, is impractical for the assessment of the massive amount of health care information on the web: for example, Google retrieved about 144,000,000 results for the query "depression treatment" submitted by the author on May 4, 2012. More importantly, this approach cannot be applied independently by information consumers who are not also subject matter experts. Therefore, researchers have developed alternative quality assessment approaches and

proposed various quality indicators to help the general public identify and access high quality health care information on the web.

Generally, the methods for quality assessment can be thought of as varying along two dimensions (shown in Table 2-1). First, depending on employed quality indicators, assessment approaches can be categorized into those that use indirect quality indicators and those that use content indicators. Second, according to the nature of rating process, the approaches can be divided into manual vs. automated processes.

Table 2-1 Quality assessment approaches and indicators

Web-based Health care information --- Quality Rating Approaches	**Indirect Indicators**	**Content Indicators**
Manual Rating	• Accessibility and availability • Author credential • Copyright notice • Currency of web pages (last update) • Disclosure of advertising • Disclosure of sponsorship • Easy of use (e.g. navigation) • Organization credentials • Readability (e.g. writing skills) • Site contact information • Site design and aesthetics	• Authority of cited medical literature • Currency of information content • Rating items in DISCERN (Charnock, 1999) • Priori items from evidence-based health care guidelines (e.g. Griffiths & Christensen, 2002)
Automated Rating	• Author information, references, currency of website, disclosure of sponsorship, advertising, copyright, etc. (e.g. Wang & Liu, 2007) • Hyperlink count • Google PageRank • Site domain name • Site traffic	• Keyword-based rating (e.g. Griffiths et al., 2005) • Semantics-based Sentence labeling (approach proposed in this study)

2.2.1 Indirect Indicators

While some early studies (e.g. Charnock et al., 1999; Berland et al., 2001) attempted to establish content-based indicators to assess information quality, most approaches have focused on developing quality indicators that do not directly reflect information content: for example, the utilization of accountability standards such as disclosure of authorship, site ownership, editorial board, etc. (e.g. Chen et al, 2000, Smith, 2002). In Fallis and Fricke's study (2002) of web pages on the treatment of fever in children, the display of a HONcode certificate and indication of copyright were both correlated with information accuracy. A study by Barnes et al. (2009) found that bipolar disorder websites with an editorial board or affiliation to a professional organization had higher quality information on bipolar disorder and its treatment than did those lacking these characteristics.

These non-content based quality indicators have also been referred to as "indirect indicators" or "proxy indicators" in previous literature (e.g. Burkell, 2004; Frické et al., 2005). In a systematic review, Eysenbach et al. (2002) analyzed the results of 79 studies evaluating the quality of health care information on the web and identified the most frequently used indirect indicators. In the studies reviewed, the 5 indicators that were most frequently used as markers of information quality were provision of references, disclosure of authorship, provision of content creation/update date, disclosure of author's credentials, and provision of email contact (Eysenbach et al., 2002).

Recently, web document usage metadata such as hyperlinks (e.g. in-link counts to a website and Google's page rank of site home page) and Internet user behavior metadata (e.g. time that a user spends on a site/page, traffic to a website, etc.) have also been used as information

quality indicators. For example, in a study on carpal tunnel syndrome websites, Frické et al. (2005) found that websites with higher information quality received on average more hyperlinks from external websites than did sites of lower quality. Also, in a quality rating study of depression websites Griffiths and Christensen (2005) found that there was a moderate correlation between the medical professionals' quality rating score and Google PageRank for the websites.

One major advantage of approaches based on indirect indicators is that these indicators are easily understandable by health care information consumers who are not necessarily subject matter experts. Burkell (2004) found that health care information users (61 participants) had significantly higher confidence in their ability to rate indirect indicators than rating content indicators. In practice, some of these indirect indicators have been included in Internet health care information quality evaluation programs, such as URAC Health Web Site Accreditation, MedCERTAIN and HONcode.

Another advantage of using indirect indicators is that the collection and processing of such indicators is relatively easy to automate. Wang and Liu (2007), for example, developed a JAVA-based automatic indicator detection tool to collect website authorship, sponsorship, copyright and other measurable indicators. In addition, website usage indicators such as in-link counts, Google PageRank, and website traffic are available from Web service providers such as Yahoo!, Google, and Alexia. These service providers provide programming interfaces to allow users to implement automated collection of such indicator values.

In spite of the above advantages, quality rating approaches based on indirect indicators have evident shortcomings. The validity of this type of approach for assessing web-based health care information quality is debatable. Evaluations of different indirect indicators (Frické & Fallis 2002; Griffiths & Christensen 2000; Martin-Facklam et al. 2003; Frické & Fallis, 2004, etc.) have suggested that these indicators bear at best a tentative relationship to information quality. In a study examining the quality of depression websites, Griffiths and Christensen (2000) demonstrated that the quality score which was calculated based on indirect indicators including site authorship, source references, disclosure and currency did not correlate with either the content quality score as measured by a rating scale developed from evidence-based clinical guidelines published by the Agency for Health Care Policy and Research (AHCPR, 1993) or with medical professionals' subjective judgment of the overall quality of sites. Fallis and Frické (2002) conducted an empirical study of websites providing information about treatment of fever in children to examine the validity of 11 different indirect indicators of information quality: the URL domain, currency, the HONcode logo, advertising, author identification and qualifications, copyright, contact information, spelling errors, exclamation points, references to peer-reviewed medical literature, and the number of in-links received by the site in question. The direct measure of information quality involved a comparison of site content with an instrument (i.e. gold standard) developed from authoritative sources on treating fever in children. The results show that only three indirect indicators correlated with the direct measure of information quality: displaying of HONcode logo, organization domain (i.e. .org) and the display of a copyright disclaimer. In addition, Frické & Fallis (2002) and Frické et al. (2005) conducted studies to test the validity of indirect indicators using websites of different health subjects, specifically cancer and carpal tunnel syndrome. In both studies,

medical experts determined the quality of the websites in question by assessing information correctness and comprehensiveness. Their results show that indirect indicators are at best inconsistently related to information quality. The number of in-links to site main page, for example, was correlated with information accuracy in the study on carpal tunnel syndromes site, but no statistically significant correlation between these variables was observed in the study on cancer websites.

The problem with indirect indicators is that they bear no necessary relationship to information quality. Frické and Fallis (2002) pointed out if a quality indicator is valid it should be difficult for authors of websites with low-quality information to display the indicator: otherwise, the utility of evaluation techniques will decrease since evaluatees can gain higher evaluations by adjusting site features to fit the indicators without improving content. In contrast to these principles, many indirect indicators can be easily satisfied by websites without changing information quality. For instance, it is not difficult for a site owner to register an ".org" domain name, which could make readers feel the site is an authoritative source. Even in-link counts and website traffic, indicators which are relatively difficult to forge, can technically be distorted by link spamming and traffic spamming. Furthermore, a variety of factors not associated with high information quality could result in a large number of external in-links to a site. According to study by Vaughan and Thelwall (2005), for example, reasons for attracting in-links to a site can include quality-irrelevant factors such as language of website, geographic factors, etc. Furthermore, in a more recent study (Khazaal et al., 2012) that evaluated the quality of websites in six different medical conditions including social phobia, bipolar disorders, etc., researchers found that although

sites holding the HON label had higher content quality scores than sites without the HON label, the difference was not statistically significant. In contrast, in the same study, a content-based indicator, i.e. the DISCERN score, was significantly associated with content quality. The researchers attributed the difference to the fact that the criteria for the HON label certification are more closely related to ethical standards than to content quality (Khazaal et al., 2012).

2.2.2 Content Indicators

Another type of quality assessment research focuses directly on website content. Crespo (2004) proposed that health care information seekers should rate online health care information against clinical guidelines and expert consensus documents. In this approach, website content is assessed with respect to indicators of content accuracy and completeness.

Compared with indirect indicators, content indicators are more reliable in providing accurate quality assessment. Among the content-based indicators or quality rating instruments, DISCERN (Charnock et al, 1999) and its variations, for example Brief DISCERN (Khazaal et al., 2009), are well-known as domain-independent instruments for rating the quality of treatment information. Both DISCERNs include many content indicators among their rating criteria (see Appendix A), asking about the web page to be rated, for example, “Does it describe how each treatment works?”, “Does it describe the benefits/risks of each treatment?” etc. The DISCERN questionnaire has been found to be a reliable instrument when used by professionals. In two independent studies (Griffiths & Christensen, 2002; Ademiluyi et al. 2003), the results generated by using DISCERN were shown to be strongly

correlated with the evaluation results provided by medical expert groups (r=0.91 and r=0.8 respectively). Moreover, DISCERN also showed good inter-rater reliability (r=0.88; p<.001) in Griffiths and Christensen (2002), in which two medical professionals used DISCERN to rate the quality of Australian depression websites.

The advantage of DISCERN is that the rating is directly based on content. Previous research has demonstrated acceptable inter-rater agreement on individual items of the instrument when used by expert health professionals and "fair" agreement among consumers (Charnock et al, 1999). In addition, the use of DISCERN is not dependent on specialist knowledge of a health condition (Charnock et al, 1999). Griffiths and Christensen (2005) demonstrated in a study using depression websites that consumer and health professional DISCERN ratings were significantly correlated (r=0.77, p<.001). The disadvantage of DISCERN, however, is that the rating process is time-consuming due to its manual nature. Hence, it is not practical to use this method to rate large volumes of health care information on the web.

Following the rise of evidence-based medicine (EBMWG, 1992), many researchers (e.g. Griffiths & Christensen, 2002, 2005; CAF & ISRCG, 2007) used evidence-based health care guidelines as content-based criteria for assessing the content quality of health websites, reasoning that health care information content of high quality should be consistent with these guidelines, which reflect the best and most up-to-date evidence-based health care practice. Evidence-based health care guidelines are normally established based on systematic reviews of scientific evidence in health care and medical literature. For example, the evidence-based guidelines published by the Agency for Health care Research and Quality (AHRQ) are U.S.

federal guidelines for clinical practice that were developed by a multidisciplinary panel and underwent extensive review by panel members including a methodologist, 28 scientific reviewers, and 73 organizations (Griffiths & Christensen, 2000). The correctness and authority of such evidence-based guidelines are widely accepted by medical experts. Although these guidelines are synoptic and do not offer health care information as detailed as that provided on health sites designed to meet the information needs of consumers, they provide a practical standard for evaluating content quality.

Griffiths and Christensen (2002; 2005) developed an evidence-based rating scale to evaluate the information quality of depression websites. Their rating scale was based on the British depression treatment guideline, *A systematic guide for the management of depression in primary care*, published by the Centre for Evidence-based Mental Health (CEBMH) at Oxford (CEBMH, 1998). Some examples of the specific criteria used in the rating scale are listed in Figure 2-1. Griffiths and Christensen establish quality rating scores by comparing web site information to these rating criteria, incrementing the score by 1 for each rating scale item reflected in website content.

Evidence-Based Rating Scale for Web Content about Depression Treatment
(Examples of rating scales used in (Griffith & Christensen, 2005))

1. Antidepressant medication is an effective treatment for major depressive disorder.
2. Antidepressants are all equally effective.
3. The effectiveness of antidepressants is around 50 to 60%.
4. Full psychosocial recovery can take several months.
5. Drop out rate is same for different antidepressants.
6. The side effect profile varies for different antidepressants.

. . .

14. Cognitive therapy can be an effective treatment for depression.
15. Cognitive behaviour therapy is at least as effective as drug treatment in mild-to-moderate depression.

. . .

20. Exercise can be effective – alone or as an adjunct to other treatments.

Figure 2-1 Evidence-based quality rating scale for evaluating depression sites

The evidence-based rating scales used in Griffiths and Christensen (2002; 2005) do not require human raters to be subject matter experts. The rating task is simply to read through a website and to verify if the web content reflects the guidelines. The quality score of a site is the number of matched evidence-based rating criteria found in the website, with a maximum value of 20. The quality assessment results were found to be highly correlated ($r=0.96$, $p<.001$) with subjective assessment results from health professionals (Griffiths & Christensen, 2002).

2.3 Automated Quality Assessment Approach

As shown in Table 2-1, approaches for health care information quality assessment can be categorized into manual and automated approaches according to the rating process. Previous studies showed that most health care information consumers do not spend much effort to ensure the information they find is of high quality (Pew Internet and American Life Project, 2006). They do not verify information sources or read disclaimers before using the information, and few users even recall the web sources from which they gather health care information (Eysenbach & Köhler, 2004). It seems unlikely, therefore, that we can rely on consumers to perform their own quality evaluations. At the same time, organizing subject experts to apply a priori quality assessment is equally likely to be ineffective, given the extremely large volume of health care content on the web and the dynamic nature of content. Thus, it is important to explore approaches to the rating of health care information quality that require less human effort. Automated rating systems are obvious candidates.

2.3.1 Previous Studies on Automated Quality Assessment

To date, there have been relatively few studies focused on the automated quality rating of health care information on the web. Those studies that have been conducted include Griffiths et al. (2005), Hawking et al. (2007), Tang et al. (2009), and Wang & Liu (2007). Wang and Liu developed an automatic indicator detection tool (AIDT) to collect indirect indicators of information quality, including, for example, the disclosure of editorial review process, date of last update, etc. Their detection tool analyzes the content of each Web page using an HTML parser and searches for indicators in metadata, HTML text, and HTML tags (Wang & Liu, 2007). They reported that the detection performance for indicators reached 93% recall (i.e.

the percentage of all indicators that were successfully detected) and 98% precision (i.e. the percentage of true indicators among the detected ones). However, their study focused only on automation of the indicator detection, and did not implement quality assessment using the detected indicators. Considering that previous studies of quality rating, as reviewed in Section 2.2.1, showed controversial results about the correlation between indirect indicators and the content quality, the ability of AIDT to automatically assess health care information quality on the web seems at best in question, if not demonstrably ineffective.

In the other three studies an automated rating approach based on keyword analysis was used to assess the quality of the health information. Griffiths et al. (2005) used a variant of the *relevance feedback* technique (Salton & Buckley, 1990; Koenemann, 1996; Dunlop, 1997) from the field of information retrieval to implement the automated quality rating of depression websites according to the evidence-based rating scale shown in Figure 2-1. The core of this approach was to train a pair of "standard" queries formed by 20 keywords and 20 two-word phrases: one query for content relevance and one for content quality. A *relevance query* was developed by contrasting the term probabilities in a set of training web pages relevant to depression with those in a set with low relevance. Human rating was used to evaluate the relevance and the quality of these training web pages. Specifically, the quality was evaluated according to the evidence-based rating scale shown in Figure 2-1. Using this technique, a standard relevance query consisting of a set of weighted terms (words and two-word phrases) was developed which had strong discriminating power to differentiate relevant pages from non-relevant ones (Griffiths et al., 2005). Similarly, a quality query was developed by contrasting high quality web pages rated according to the evidence-based

guideline with a set with lower quality by this measure. Once the training was completed, the queries were run against testing websites. For a given website, the *relevance score* and *quality score* were calculated using a linear function based on the similarity between the web pages contained in the site and the "standard" queries. The final quality rating score of a website was a linear combination of two scores. In Griffiths' study, the researchers used 30 depression testing websites to verify the performance of the automated quality assessment system. They found high correlation ($r=0.85$, $p<0.001$, $n=30$) between automated quality rating scores and manual quality rating scores measured according to evidence-based rating criteria in Figure2-1. In contrast, the correlation between Google PageRank and manual quality rating scores was only 0.23 ($r=0.23$, $p=.22$, $n=30$).

The keyword-based approach developed by Griffiths and her colleagues provided an automated rating solution. Using the automated solution, it is possible to evaluate millions of depression websites. No effort to evaluate quality is required on the part of the information user. The quality rating results, which are normalized quality scores that indicate the quality of the depression sites, allow health care information users to identify high quality health information sources. In contrast to approaches using indirect indictors, the rating approach developed by Griffiths and Christensen relies on text content (i.e. keywords and phrases) instead of accountability or other types of metadata. Given the evidence, presented above, the content-based approach is likely to lead to better quality ratings.

However, the weakness of the keyword-based automated rating approach (Griffiths et al., 2005) is that the meaning of text content is not utilized for quality rating. Instead, the rating

methodology transforms the depression web page into a vector of keywords and relies on its similarity to the trained standard queries to predict the information quality. Specifically, the keyword vectors are formed by only the 20 highest discriminating keywords (Griffiths et al. 2005), and thus can hardly represent the full content of a website in a comprehensive way. In addition, the semantic connections between keywords are ignored in this solution. Therefore, although the quality rating is conducted based on text content, the actual meaning of the text of the health care information is not exploited for quality rating in the keyword approach.

2.3.2 Other Related Techniques

Other techniques that are potentially useful for automated content-based quality assessments include natural language processing and machine learning methods. Few published works in quality assessment have employed these methods. The effectiveness of these techniques for health care information quality rating is explored in this study.

Natural language processing and related techniques are applied in this study to implement semantic parsing and processing. Through semantic parsing, sentences expressed in natural human language can be mapped to a formal representation of semantic concepts and relationships. Thus, computer programs can be developed to rate the health information quality based on the semantics of text in web pages. Studies on developing information extraction techniques have attracted great interest from researchers in multiple disciplines, mainly computer science. Extensive research activities started in nineteen-eighties, including the Message Understanding Conferences initiated in 1987 and financed by DARPA (Grishman & Sundheim, 1996), and later the DUC (Document Understanding Conferences)

and TREC (Text Retrieval Conferences) run by National Institute of Standards and Technology (NIST). Techniques such as name entity recognition have proven effective for dealing with text in diverse domains, including the biomedical domain (e.g. Nadeau & Sekine, 2007).

The technology for semantic processing and analysis is in a growth stage. Although there is as yet no universal tool that can in general solve domain-independent text understanding questions, semantic analysis and processing has been successful in some domain-specific applications. Particularly in the health care information knowledge domain, for example, semantic parsing and processing has been successfully used in biomedical concept annotation (e.g. CONANN) and summarizing biomedical text (Reeve, 2007), in classifying medical patient records (Chen et al., 2010), and in extracting medication information from text clinical records (Deléger et al., 2010). Given the success of these studies, it is worthwhile to attempt automated quality assessment of the web health care information based on shallow semantic analysis of text, so that the content of web documents and the rating criteria can be compared through semantics.

In the health care knowledge domain, many tools are available to facilitate text processing functionalities including morphological processing, syntactic processing, grouping synonyms terms into concepts, etc. There are also controlled vocabulary resources available such as MeSH published by the National Library of Medicine. In addition, the SNOMED CT maintained by the International Health Terminology Standards Development Organization (IHTSDO) provides a comprehensive clinical terminology for clinical terms. In particular,

the U.S. National Library of Medicine provides a tool called Unified Medical Language System (UMLS). The UMLS integrates more than 60 families of controlled biomedical vocabularies including MeSH and SNOMED CT. It was developed to reduce the barriers to effective retrieval of machine-readable information by including the variety of ways the same concepts are expressed in different sources and by different people, and also to enable the representation and distribution of health care information among systems (Humphreys et, al., 1998). With these two advantages, UMLS can be an excellent infrastructure for the effective transformation of a great variety of text in biomedical domain into normalized semantic annotations. Therefore, the UMLS tool is used in this study to generate the semantic representation of web health care content written in English natural language and rating criteria in order that they can be compared.

Another potentially useful technique for quality assessment is text classification and related machine learning algorithms. In our semantics-based quality assessment approach, quality scores are assigned based on semantically comparing the text with the evidence-based quality rating criteria. This study tried to implement the comparison through text classification, in which text can be classified into predefined classes according to content. A number of statistical and machine learning techniques have been developed for text classification, including rule-based decision system, Naïve Bayes, support vector machines, and maximum entropy models (Sebastiani, 2002). Many early text classifiers were based on keywords extracted from the documents, with the assumption that a keyword is a unique representative of a distinctive concept or semantic unit. Thus, these earlier classifiers do not include strategies for effectively handling the polysemy and synonymy observed in natural language:

a word may represent multiple different meanings, and people can choose different words to refer to the same meaning. However, in some text classification studies (e.g. Wiener et al., 1995; Liu et al., 2004; Wang et al., 2005; Wang & Liu, 2009) natural language processing techniques (latent semantic analysis or LSA) proved to partially resolve the problems of word choice and redundant semantic relationships in text classification. This technique analyzes the associative semantic relationships between a set of documents and the terms they contain by constituting a latent semantic space related to documents and terms. A hint from the LSA studies is that the utilization of semantics will likely be beneficial to text classification in this study and hence it is worthwhile to attempt text classification in combination with semantic parsing and analysis.

Text classification has been successfully used in various domains to solve different application problems, such as e-mail spam filtering (e.g. Sahami et al., 1998), categorizing news articles into topics (Schapire & Singer, 2000), and assigning international clinical codes to patient clinical records (Chen et al., 2010). This thesis explores the application of text classification to a new application area – i.e., rating the content quality of health documents on the web.

2.4 Research Goals of This Study

With inspiration from the work of Griffiths and her colleagues, the current study aims to provide health care information quality ratings through an automated analysis of the semantics of information content. Two parallel quality rating systems are developed and evaluated in this study: a rule-based system and the prototype of a machine learning system.

Both use natural language processing to analyze the semantics of the health care information text and based on the common semantic analysis results compare the text content against rating criteria derived from evidence-based clinical guidelines. The difference between the two systems is the method used for identifying matches between the web text and the rating criteria. In the rule-based quality rating, the matching of semantics is conducted using patterns extracted through manual knowledge engineering; in machine learning based rating, the computer system learns patterns by itself from semantic parsing results of training texts. As will be illustrated in Chapter 5 and 6, the quality rating results generated by both systems are comparable to human rating results, demonstrating that a semantics-based quality rating approach is promising in providing practical assistance to health care information consumers in identifying high quality health care information on the web.

To the best of our knowledge, the information quality rating approach in this study is new in that it tries to apply the analysis of text semantics to implement quality assessment. The research questions to be addressed include:

1) For the purpose of content-based quality assessment, how can the semantic representation be constructed so that the text content of the web health care document can be conveyed effectively? The semantic representation generated certainly needs to be computer readable and processable. More critically, it needs to capture and represent sufficient and necessary semantics of the text in order to enable a computer system to compare the content of a web document with the quality rating criteria.

2) Based on the created semantic representation, how can a computer application be built to identify successfully whether the content of a web health care document is in concordance with a rating criterion? Technically, will the rule-based and machine learning based classifications proposed in this study be effective approaches? As quality rating results rely on the degree of concordance between the text content and the rating criteria, accurate identification is important for the implementation of automatic quality rating.

In order to keep the research questions to a manageable size, the scope of this study is limited to the treatment information for a single medical condition, namely depression. The Unified Medical Language System (UMLS) and related natural language processing tools are employed to process text on the health care web pages. It should be noted that these resources are useful across biomedical domains and therefore the limited subject matter scope in this study should not be taken as a limit on the applicability of the results.

Chapter 3
Research Methodology Part I – Experiment Design

This study presents a new approach to rating the quality of health care information on the web. In contrast to methods using keywords term frequency, accountability metadata, or other information quality indicators, the current study approaches information quality assessment through training computer programs to rate content quality based on the semantics of sentences in a web text.

3.1 Overview of Semantics-based Quality Assessment

Generally speaking, the quality assessment approach used in the current study is to train computer programs to complete quality ratings in a manner analogous to that used by human beings: by contrasting the semantic content of the online text with a 'gold standard' of clinical evidence. This study follows previous research (e.g., Griffiths & Christensen, 2002; 2005) in taking depression as the information domain, and in using established clinical guidelines as the standard against which information is evaluated. The established clinical guidelines for depression contain twenty evidence-based rules covering different aspects of depression treatment (Griffiths & Christensen, 2005). The current study evaluated the quality of online health care information about the treatment of depression by using an automated process to compare online information to rating criteria based on these established clinical guidelines. The quality score assigned to each web page was the number of unique rating criteria identified in the texts. The approach was to train the computer programs to analyze the semantics of sentences in the web pages. If the content of a sentence was identified to be

in concordance with a rating criterion, the quality score for the web page was incremented by one. However, multiple identification of the same criterion in a web page did not further increase the quality score.

The following steps were involved in the webpage quality rating study:

- Preparing the data set, i.e. collecting a group of web pages addressing depression treatment.
- Establishing quality rating criteria based on the evidence based depression treatment guidelines (more details in Section 3.3 and 3.4).
- Organizing human raters to conduct quality rating on all collected web pages by using the evidence based rating criteria (details available in Section 3.5).
- Splitting data set into two mutually exclusive parts: training set and testing set.
- Developing a semantic processing tool to generate shallow semantic representation of health care text collected in above steps (more details in Section 4.1 and 4.2).
- Developing a classification tool which relies on shallow semantic representation to classify sentences according to sentence content (more details in Section 4.3); using training web pages to train the application tool.
- Using the learned tool to classify the sentences in testing web pages based on their shallow semantic representation, i.e. identifying the sentences that are in concordance with the rating criteria derived from the evidence-based clinical guidelines. Assigning quality score to each web page based on the number of unique rating criteria identified in the page content.

- Evaluating the quality rating performance by quantitatively comparing the quality scores rated by the automatic quality rating tool with scores rated by human raters (Evaluation results available in Chapter 5).

3.2 Data Sampling

The corpus for this study comprised a total of 201 web pages on the topic of depression treatment. The sample data were obtained from multiple sources, as listed below, in May 2009. Thirty-one pages were selected from the corpus using stratified random sampling to serve as testing data, leaving 170 pages as training data. The generation of data corpus is introduced below.

Previous studies on online health care information-seeking behavior (Graham et al., 2006; Pew Internet and American Life Project, 2006) show that users typically access health web pages directly from Web-based search engines (66%) or from consumer health sites/portals (27%). Since the purpose of this research is to explore a new approach for automated information quality rating in order to improve users' health care information practice on the Internet, this study examines the system performance using web pages that would likely be encountered by consumers in real life experience. Therefore, the web pages that comprise the dataset were collected from Web search engines and health care portals. The URLs of these web sources are listed in Appendix B. The data collection details are provided below.

3.2.1 Data Sample Source - Search Engines

Ten Web search engines were used to retrieve candidate web pages. Five were common search engines:

- Google
- Yahoo! Search
- Microsoft Bing Search
- Ask.com
- AOL

The other five were medical search engines:

- OmniMedicalSearch
- HealthFinder
- HealthLine
- MedNar
- WebMD

Google, Yahoo, Bing, Ask.com and AOL were selected because they are the most popular Big Five (Nielsen, 2010; comScore, 2011). Their shares of the U.S. explicit core search market in descending order were 65.5%, 15.9%, 14.3%, 2.9% and 1.4% as of June 2011 (comScore, 2011). Medical search engines were also used to collect sampling data because online users also tend to seek health care information using medical search engines since these search engines are focused on health related topics and tend to include only credible sources (e.g. Mednar, 2009). Although there was no research examining the market share of the medical search engines, the five engines used in this study are all famous and long-standing portals in this area (Leman 2008; About, 2009;). Among them, OmniMedicalSearch and HealthFinder.gov comply with the HONcode standard (Boyer & Geissbuhler, 2005; Baujard et al. 2011) for providing trustworthy health information.

Against each of these ten search engines, a two term search query [q = depression treatment] was submitted to retrieve web pages. URL candidates were collected from the first three pages of retrieved results (10 returned URLs per page, making it a total of 30 URLs per

search engine), because most online users rarely go beyond the first three pages of returned results (eWebMarketing, 2009). Each URL for the 300 items was then examined in order to filter out invalid candidates: for example, inaccessible URLs, duplicate URLs, or those that were inappropriate for any other reason. The details of filtering process and sampling criteria are listed in section 3.2.3.

3.2.2 Secondary Data Sample Source – Health Care Web Portals

Health care web portals were another credible source for collecting web page samples. Four health care web portals in English language were used in this study: namely Medline Plus in United States, HealthlinkBC in Canada, HealthInsite in Australia, and National Health Service (NHS) in United Kingdom.

Medline Plus was chosen because it is one of the most popular health web portals, providing easy-to-understand information for common health care information consumers. It provides authoritative health care information from the world's largest medical library NLM (the National Institutes of Health), and other government agencies and health-related organizations (Medline Plus, 2011).

The Medline Plus home page for depression is http://www.nlm.nih.gov/medlineplus/depression.html and online users navigate from here to reach other hyperlinks of their interest. The hyperlinks on this portal page are grouped by subtopics. Since the topic for this study is the treatment of depression, URL candidates were collected from treatment related subtopics only. Then candidate pages were manually

examined to guarantee that content was within the depression treatment scope. The hyperlink navigation depth was limited to two jumps.

With the same approach, sample web pages were also collected from a Canadian based health care web portal www.healthlinkbc.ca, an Australian site www.healthinsite.gov.au, and a national health service website in UK, http://www.nhs.uk. Each of these sites is hosted by the government or a governmental agency, and the sites are committed to high editorial and ethical standards in the provision of content and related services (HealthLinkBC, 2011; HealthInsite, 2011; NHS, 2011).

3.2.3 Filtering Out Invalid Web Page Samples

URL candidates collected from search engines and health care web portals were pooled together. The resulting URLs were filtered to remove duplicates – if multiple candidates had the same page content, only one page was included into data corpus. Pages which did not have relevant content were also filtered out. For example, if the text was about depression diagnosis instead of treatment then it was removed. In addition, while audio, video and image web pages, and web pages including tables and graphs can have relevant content, they were also excluded from this study because the proposed quality rating approach is limited to dealing with text format web content. Below is a detailed list of reasons for excluding certain web pages from final samples:

- pages which focus on other diseases instead of depression, or pages that address depression, but discuss only non-treatment topics such as diagnosis; --- *determination was based on document heading & sub-heading*.
- pages protected by password.

- pages not in text format (e.g. video/audio clips, PPT slides).
- pages with tables or spread sheets as major part of page content.
- portal pages which do not have their own content, but just hyperlinks referring to other relevant pages. (e.g. URL menus/categories, list of search returns from search engines)
- pages which have article titles or bibliographic information only
- home pages of business or organizations (e.g. medical center or depression clinic)
- pages too long for human rating (e.g. online books or chapters) --- *they were filtered out due to the consideration of human rating expense*.
- advertisement pages which do not really provide depression treatment content, such as Amazon book advertisement
- academic articles which are targeted for professional audience instead of public online users --- *due to academic complexity, some very specific research questions and terminologies can make the articles not quite understandable for most common users and human raters to conduct rating*.

After the filtering process, the final corpus contained 201 valid web pages. The URLs of all 201 sample pages are saved in Appendix B. Due to the volatile nature of online content, it was anticipated that the web documents could experience content change from time to time or the document could become inaccessible. Therefore, the content of each sample page used in this study was saved into a Microsoft Word document so that researcher and human raters have consistent dataset for analysis.

3.3 Gold Standard for Quality Rating

Evidence-based medicine practice has been advocated in everyday care since the original model of evidence-based medicine was presented in 1992 in the *Journal of the American Medical Association* (EBMWG, 1992). Many evidence-based clinical practice guidelines have been established under the sponsorship of governmental agencies such as the Agency for Healthcare Research and Quality (AHRQ, 2011) in the United States. The guidelines are normally established based on the systematic review of scientific evidence in health care and

medical literature by multidisciplinary panel including methodologists, medical experts and scientific reviewers. Researchers have experimented with the use of such evidence-based health care guidelines as gold standards for assessing the quality of health websites, and this approach has proven to be successful (e.g. Griffiths & Christensen, 2002, 2005; CAF & ISRCG, 2007). One advantage of using evidence-based health care guidelines as a rating standard is that the rating process is relatively immune to the subjective bias that might affect less structured quality evaluations. In addition, evaluation relative to these clinical guidelines does not necessarily require raters to be domain knowledge expert since most evidence-based guidelines are clear and straightforward.

Griffiths and Christensen (2002) adopted a set of evidence-based depression treatment rules published by the Centre for Evidence-based Mental Health at Oxford (CEBMH, 1998) as the quality rating standard in their study. In their study, human raters used this standard to rate the quality of depression websites. The quality of a website was measured by the number of different treatment rules reflected in the website content. The larger the number, the higher the quality score a site was assigned. This study proved that the rating scores generated using evidence-based treatment guidelines were highly correlated ($r=0.96$, $p<.001$) with the quality scores of subjective rating completed by health professionals (Griffiths & Christensen, 2002).

Following Griffiths and Christensen (2005), the current study also used the 20-item evidence-based depression treatment rules as the gold standard for quality rating. As previous researchers have made advances in efforts to summarize and refine the evidence-based treatment guidelines into a set of one-sentence statements, each treatment rule could be easily

converted into a rating criterion in the current study. Appendix C lists these twenty evidence rules. These guidelines were created in 1998, it is possible that they may not reflect the latest scientific findings in depression treatment. However, since the focus of this study is to explore whether computer program can automatically identify the sentences in depression treatment web pages that are in concordance with evidence-based health care guidelines and rate quality accordingly, the use of these guidelines in this study is appropriate as long as human raters and computer programs use them as common rating criteria.

3.4 Rating Criteria

Rating criteria were established based on the evidence-based treatment guidelines. In order to make the criteria explicit and easy for human raters to follow and thereby to minimize any inconsistency between raters, the following transformation processes were applied to the guidelines to decrease the rating criteria's semantic complexity and ambiguity.

Depending on semantic complexity, an evidence-based guideline item could be converted into one or multiple rating criteria. In this study, the semantic complexity was determined by the number of semantic propositions contained in the guideline. As shown in Appendix C, most guideline items have only one semantic proposition, such as guideline items from 1 to 7. They were used as rating criteria without any modification. Guideline items containing multiple semantic propositions were split into multiple rating criteria. For instance, the guideline #12 -"abrupt cessation of antidepressant can cause discontinuation syndrome and that antidepressants should not be stopped suddenly"- is a complex sentence and it consists of the following two "meaning pieces" which are not quite the same as each other.

1) Antidepressant should not be stopped suddenly.
2) Abrupt cessation of antidepressant can cause discontinuation syndrome.

Since it is certainly possible that one point is mentioned in a web page while the other is missing, guidelines like this can potentially cause discrepancy among human raters. To avoid such problems, criteria 12-A and 12-B were generated to correspond to guideline #12. After these transformations, the 20 evidence based clinical guidelines were converted into 23 rating criteria. The rating criteria (e.g. 12-A and 12-B) each contribute separately to the quality score assigned to a web page. That is, for each criterion represented in a web page, 1 is added to the quality score, with the repeated items counted once only. The quality score of a web page about depression treatment therefore could therefore range from 0 to 23.

In natural language, different authors can express the same semantic content in different ways. Even when two sentences express content in concordance with each other, they will in most cases differ in terms of specific semantic components (e.g., modifiers). For this reason, inter-rater rating discrepancy could happen when different raters require different levels of conformity to identify a match between a criterion and a candidate sentence. Also, intra-rater discrepancies could exist because a rater's application of the criteria could vary during the rating. For example, guideline #11 states "Once improved continuation treatment at the same dose for at least 4-6 months should be considered." The main point is about continuation of treatment after improvement. But in order to get an exact match, both "same dose" and "4-6 month" have to be covered, and even "at least" if the maximum degree of equality of content is required. For quality rating purposes, a match can be claimed when the main point of a sentence is in concordance with a rating criterion, rather than requiring a match at the finest

semantic granularity. Thus, in this study, one important principle for human rating is to focus on detecting the main point of a criterion. The matching of modifiers is nice to have, but not mandatory. For example, the main point of guideline #11 was "Once improved continuation treatment should be considered." Through this transformation a common standard for content matching was introduced to the human raters. In addition to guideline #11, such transformation was also applied to guideline #9, 13, 15, 19 and 20. Appendix C lists all the transformed rating criteria in this study.

3.5 Human Rating

Human raters were hired to evaluate the information quality of web pages in both the training and testing data sets. It was demonstrated in previous studies (Griffiths & Christensen, 2002, 2005; Griffiths et al., 2005) that the rating scale derived from the CEBMH depression treatment guideline is reliable and the inter-rater reliability is very high ($r=0.93$, $p<0.01$). Thus, it was reasoned that two raters were enough for completing human rating in this study. Rater A was a medical professional and had eight years of professional experience as a pharmacology instructor. Rater B was not a medical professional, but had one year of part-time rating experience working in the Google Quality Rating program.

3.5.1 Human Rating Process

In order to have consistent evaluation of all web documents, the following strategy was designed for the human rating process.

- A five-hour rating workshop was held for human raters to learn the evidence-based rating criteria and to exercise page rating independently. Then they exchanged and discussed rating results.

- All 201 pages were rated, before being split into training and testing data sets.
- Human raters were given the page content saved in Word document, instead of URL so that the rating results are free of impact by website name, authorship, page aesthetic design, or other factors which may cause subjective bias.
- Raters were required to label the sentence or sentence group that they identified to be in concordance with the rating criteria. A tag pair was used to include the identified sentence. For example, <Criterion 1> *sentence content* … </Criterion 1>.
- Within the labeled sentences, raters were required to underline the key words or phrases which they considered as essential semantic elements for identifying the "*criterion-like*" sentences (also referred as "positive cases" in the following chapters).
- Raters were required to complete rating independently.
- After independent rating, two human raters exchanged rating results, discussed discrepancy and reached agreement. When an agreement could not be reached, researcher reviewed the data and then made a final decision.
- The quality score for each web page was computed as the total number of distinct criteria represented on the page.

A complete version of the rating code for human raters can be found in Appendix D. It includes both the general rating guide mentioned above as well as the detailed instructions on individual cases. After two human raters finished independent rating, an intra-class correlation coefficient (ICC) testing was used to estimate the inter-rater reliability of human rating across 23 criteria in this study. The single measure ICC value i.e. ICC(3,1) was .990, with the 95% confidence interval between .979 and .995.

3.6 Training Data and Testing Data

The data corpus comprises of 201 web documents. They were assigned unique IDs from 1 to 201. Appendix B gives the URLs of these pages. The quality of each page was rated by the human raters using the rating criteria identified above. The scores ranged from 0 to 8. The quality scores of the collected pages have a skewed distribution (see Appendix B). Based on the human rating results, the web sites were pooled into 5 bands (quality ratings of 0, 1-2, 3-4, 5-6 and 7-8). Stratified random sampling was used to sample 31 testing web pages in order

to avoid generating a spuriously low correlation due to restricted range effects. The URLs of these 31 testing pages are also listed in Appendix B. The rest of web pages were used as training set to develop computer system's capability to identify sentences that are in concordance with the rating criteria.

Chapter 4
Research Methodology Part II – Auto Quality Rating Method

In the previous chapter, the experimental design was introduced, including preparation of the rating criteria, data samples and the human rating. This chapter focuses on the development of the semantics-based quality rating approach. It explains the process whereby the automatic quality rating programs generate shallow semantic representations for sentences in the web pages, and then identify among the testing web pages the sentences that are in concordance with the rating criteria through matching the shallow semantic representation of a sentence with the patterns learned from similar representation of sentences in training web pages.

4.1 Overview of Semantics-based Quality Rating

In this study, the way that the computer programs conduct page quality rating is similar to the human rating process. Humans read through a document sentence by sentence, and identify whether each sentence matches semantically with any rating criteria. If a match with a rating criterion is found, and no previous match has been identified in the web page, then 1 is added to the quality score. Given this description, the automated quality assessment in this study therefore takes a three-step approach. The first step is to convert the text of web pages into a shallow semantic representation of each sentence. The second step is to use the generated shallow semantic representations to identify the sentences of which the content is in concordance with the rating criteria. The training data set is used to develop the identification capability. The third step is to assign and calculate the quality score based on the identified matching criteria. Figure 4-1 provides an overview of the semantics-based quality rating procedure.

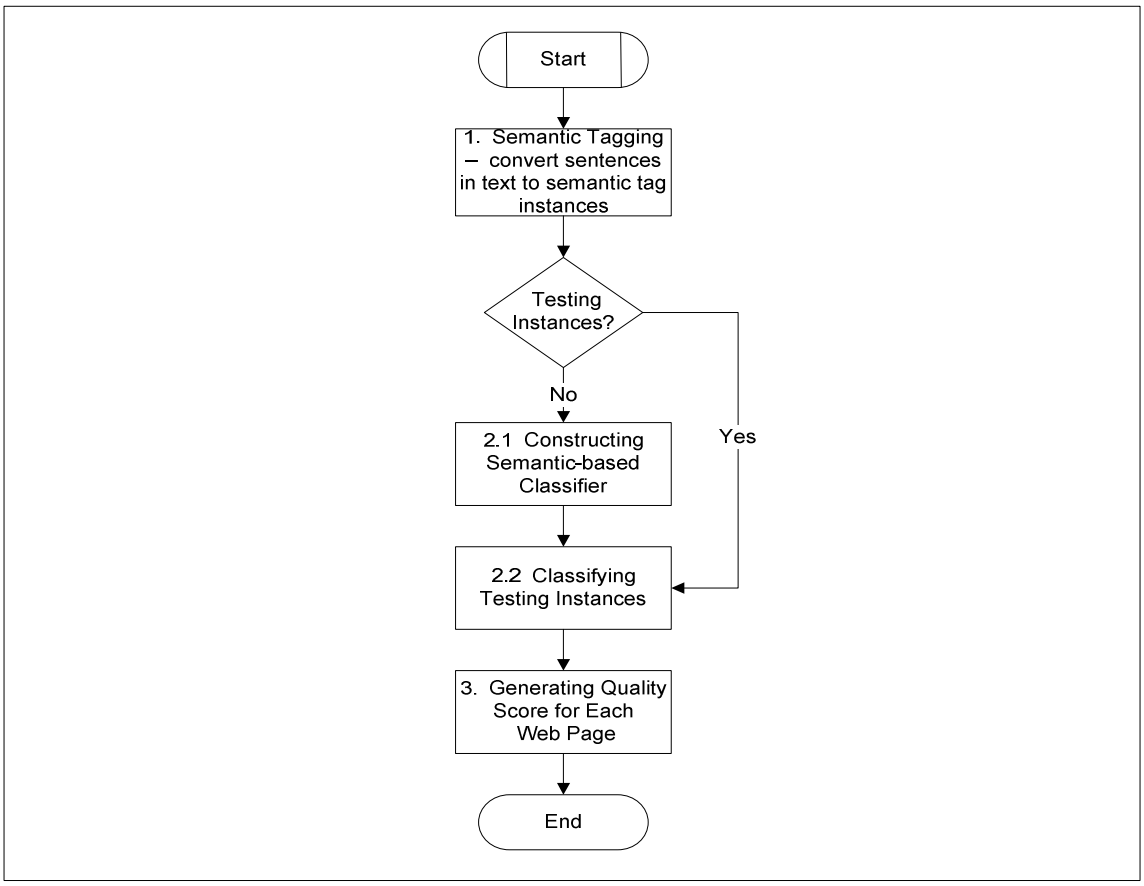

Figure 4-1 Flowchart of the semantics-based quality rating

Clearly, the task of identifying whether a sentence is in concordance with a rating criteria can be transformed into a semantics-based classification question. It is a common and well-proven ability of automated classifiers to classify text documents into different groups (e.g. according to topics) in traditional text classification studies (e.g. Han & Karypis, 2000; Kim et. al., 2005; Guan et. al., 2009). In the current research context, sentences instead of the whole document are the objects being classified. Computer programs have been designed to classify sentences into binary groups (i.e. TRUE or FALSE) with reference to each rating criterion in Appendix C. It is called semantics-based classification because the classification

relies on semantic components of the sentences and the classification result indicates whether the content of a sentence is in concordance with the rating criterion being examined.

4.2 Semantic Tagging

The semantics-based sentence classification task depends on the semantic representation of sentences. In this study, semantically important units in sentences, particularly notional words and phrases, are semantically tagged, since these units are essential components of the meaning of a sentence. Sentence classification has been implemented based on this semantic tagging.

4.2.1 Semantic Representation of Sentences

In this study, the purpose of semantic tagging is to identify and represent semantically important components of depression treatment statements. The semantic representation of a sentence deals, in part, with concepts and relationships between concepts (as shown in Figure 4-2).

A positive sentence of rating criterion #1:
"**Antidepressant medication can be effective in treating major depressive disorder**."

↓

Semantic representation:
Antidepressants -- Treat -- Depression

Figure 4-2 Semantic representation of a positive sentence of rating criterion #1

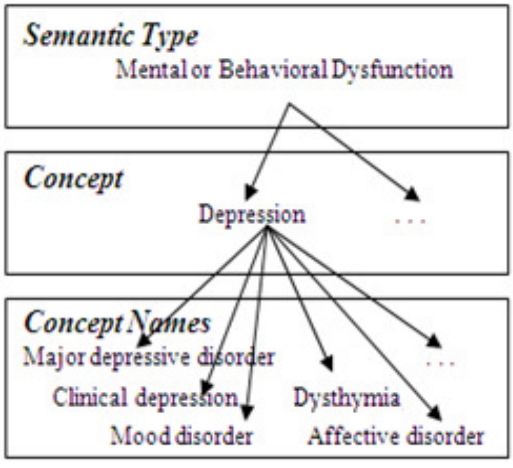

Figure 4-3 Semantic mapping of concepts

Semantic concepts play an important role in interpreting a sentence. The semantic representation of a sentence can contain one or multiple semantic concepts. Since every depression treatment criterion has a unique theme, it contains a unique set of semantic concepts. The sets of concepts can be used to differentiate the criteria.

In the training texts, the human raters found instances of sentences that have meaning consistent with the rating criteria. Such instances and their nearby sentences are an important source from which to extract theme-related semantic concepts. In addition, these sentences in the training data set are also important source from which the computer can learn how the semantic concepts are connected with each other for expressing a specific rating criterion. Due to the nature of natural language, there can be more than one way to express the same meaning. These differences among expression variations can be reflected in the sets of semantic concepts being used and also the way in which a set of concepts are assembled together. It is anticipated that such patterns to be identified in the testing text do likely show up in training text as well because statistically the usage frequency of different expressions is not supposed to have statistically significant difference between their distributions in the training and testing data sets. Therefore, identifying whether a sentence in the testing pages is in concordance with a rating criterion can be implemented through checking whether its semantic representation has a pattern in common with one of the training sentences that human raters marked as positive cases of the specified rating criterion. Although it is not necessary that every expression variation in the testing pages be covered by patterns learned from the training set, relatively frequently used semantic representation patterns will more

likely be learned during training as long as the training set size is large enough. Therefore, through training the learned computer system is expected to be able to cope with the majority of, if not all, expression variations occurring in the testing pages.

4.2.2 Text Processing

Before any other text processing occurs, the text is cleaned to remove formatting factors that could cause processing failures or erroneous results. For example, without this text cleaning step being included, citation references in a web page— "`research shows it is good.5,76,211 Antidepressant . . .`" —caused sentence splitting or tagging errors; certain characters such as "[" and "]"also caused syntax errors in tagging results. After cleaning, the text is ready for semantic tagging which converts each sentence in the text into semantic representations as mentioned above. The semantic tagging process, described in the next section, has been programmed in JAVA code.

4.2.3 Semantic Concept Tagging

A central theme in this study is the conversion of a natural language sentence into a shallow semantic representation consisting of semantic concept tags. In this study, natural language processing and semantic processing tools provided by the National Library of Medicine (NLM) are used to generate these semantic concept tags.

The Unified Medical Language System (UMLS) knowledge sources built by NLM is a compilation of more than 60 controlled vocabularies in the biomedical domain including Medical Subject Headings (MeSH), Systematized Nomenclature of Medicine (SNOMED),

Physicians' Current Procedural Terminology (CPT), Clinical Terms Version 3 (Read Codes), etc. The UMLS provides semantic concepts, hierarchical structures and relationships for each vocabulary. For instance, the relations in Figure 4-2 are described in the UMLS semantic network; semantic types, concepts, and concept names in Figure 4-3 can be found in the UMLS Metathesaurus. The 2009 version UMLS Metathesaurus is comprised of over 2 million biomedical concepts and 9 million concept names, each of which has variant terms with synonymous meaning (UMLS, 2009). Thus, this source is used in this study to provide a comprehensive vocabulary for depression treatment related terminologies which are frequently used in written documents. The UMLS semantic network gives a hierarchy to determine the most specific semantic type to be assigned to a Metathesaurus concept. The levels of the hierarchy from bottom to top are concept names, concepts, and semantic types (as shown in Figure 4-3). In this study, the hierarchical links assist in converting synonymous concept names to unified semantic concepts and types, which are like controlled vocabularies used in standardized classification systems.

The Semantic Knowledge Representation (SKR) project at the NLM developed the SPECIALIST NLP system (McCray et al. 1994; National Library of Medicine, 2009). It provides a framework, i.e. the SPECIALIST Lexicon and associated lexical variant programs, to support syntactic analysis and semantic interpretation of biomedical text. The SPECIALIST Lexicon contains more than 140,000 entries of general and medical terms about English verbs, nouns, adjectives and adverbs (Fridsma & Aronson, 2002).

The MetaMap program developed by SKR provides an API to integrate the above resources and tools. In this study, JAVA programs have been written to map noun phrases in text to concepts in the UMLS Metathesaurus using the MetaMap JAVA API (version 2.6). Paragraphs are successfully split into sentences; then nouns and noun phrases in sentences are labeled with their semantic tags derived from controlled vocabularies. Figure 4-4 gives some examples of semantic concept tagging results (i.e. MMTx tagging results).

URL --- http://www.answerstodepression.com/whatisdepression/whatisdepression.php

Free text 1 --- "What is Depression?"

MMTx Tagging Result ---
```
[0, 0, 20, [2, Depression, 8, 18, 1000, [[1000, Depressive disorder,
8, 17, Mental or Behavioral Dysfunction]]]]
```

Free text 2--- "People with a depressive illness cannot merely "pull themselves together" and get better."

MMTx Tagging Result ---
```
[8, 520, 610, [63, People, 520, 525, 1000, [[1000, Persons, 520, 525,
Population Group]]], [64, depressive illness, 527, 551, 1000, [[1000,
Depressive disorder, 534, 543, Mental or Behavioral Dysfunction]]]]
```

Figure 4-4 Semantic tagging result of free text

The labeled results are comprised of the following information: The first field is the sequence ID of the sentence in the text, 0 and 8 in the first and the second instances, respectively. The second and third fields are the beginning and ending offsets of the sentence, 0 to 20 and 520 to 610, respectively. The noun phrases in each sentence are tagged. In the first instance, "Depression" is annotated, with the token sequence ID being 2 in the text. The phrase corresponding to this tag is "Depression?" with position ranging from 8 to 18, including the

question mark. The UMLS semantic concept assigned to this tag is “Depressive disorder”, due to the core term “Depression”, which starts from position 8 and ends at 17. In the second instance, the first two tags are “people” and “depressive illness”. The token sequence ID of the tag “people” is 63. The second tag, “depressive illness”, corresponds to a prepositional phrase “with a depressive illness” which is from position 527 to 551. 64 is the sequence ID of the token “with”, from which the prepositional phrase starts. The token “depressive”, which the MMTx tagging tool considers as a core term for determining the semantic concept, starts from position 534 and ends at 543. Although different text terms were used in these two instances, the tagged semantic concept in the second instance was same as in the first, i.e. “Depressive disorder”. The mapping score was located in the field after the ending offset of the noun phrase head. It indicated the confidence of mapping the text to the specific Metathesaurus concept. The mapping score ranged from 0 to 1000 according to MetaTag’s score scale.

The examples in Figure 4-4 illustrate that tagging using the MetaMap API can cope well with part-of-speech (POS) processing. In addition to lemmatization of biomedical nouns and noun phrases, the MetaMap API also supports the normalization of different names for the same semantic entity, even including acronyms and synonyms. This functionality is important since there are a large number of expression variations for medical terms in the depression treatment domain such as “depression”, “antidepressants”, “medication”, “treatment”, etc. Also shown in Figure 4-4, the MMTx tagging result is generated based on the phrase head, hence lexical items with relatively less semantic value such as prepositions, determiners, conjunctions and punctuations are not included in the tagging results.

4.2.4 Lexical Tagging

In addition to using MetaMap to process nouns and noun phrases, two other SKR tools, TaggerClient (v2.4.c) and Lexical Variant Generator (LVG), are used in this study to process verbs, adjectives, adverbs, etc.

TaggerClient is a tokenizer. It is used to tokenize text. The output of TaggerClient contains not only the token per se, but also important metadata including the beginning and ending offsets of the token, the part of speech, and the sequence number of the token inside the text. Furthermore, lemmatization is applied by using the LVG tool to reduce lexical variations of different types, including inflections and conjugations, word order in multi-word terms, alphabetic case, punctuation, and possessives. For instance, “ceases” is transformed to “cease”. An advantage of the LVG tool is that it consists of a collection of independent submodules which can be combined in any way in order to generate the variants desired by the user (Aronson, 1994). The submodules include lowercase/uppercase processing, removal of genitives, removal of punctuation, generation of inflectional and derivational variants, and other natural language processing functions. In addition to string normalization, the LVG tool also includes a synonym processing submodule. It provides a built-in synonym dictionary that stores synonym pairs. For example, a pre-defined pair “cease – stop” makes it possible to convert the expression “ceases” to a unified semantic tag “stop”. Moreover, the LVG tool provides an interface that allows users to supplement the synonym pair definitions, if needed. Figure 4-5 shows an example of LVG processing output. Sequentially, the tagging result

includes beginning offset, ending offset, original token, POS, synonym, and the token's sequence number within the hosting sentence.

```
|0|3|What|pron|what|0
|5|6|is|aux|am|1
|8|17|Depression|noun|depressive disorder|2
|18|18|?|punctuation|NULL|2
. . .
|520|525|People|noun|people|0
|527|530|with|prep|NULL|1
|532|532|a|det|a|2
|534|543|depressive|adj|depressive|3
|545|551|illness|noun|disease|4
|553|558|cannot|modal|can|5
|560|565|merely|adv|merely|6
|568|568|"|punctuation|NULL|6
|569|572|pull|adv|pull|7
|574|583|themselves|pron|themselves|8
|585|592|together|adv|together|9
|593|593|"|punctuation|NULL|9
|595|597|and|conj|NULL|10
|599|601|get|verb|get|11
|603|608|better|adj|best|12
|609|609|.|punctuation|NULL|12
|611|617|Without|prep|without|0
. . .
```

Figure 4-5 LVG tagging result example

As shown in the LVG tagging results, every token is processed as a unit, in a way that is different from MMTx tagging in which a unit is a noun phrase. To capture all semantically important components for conducting semantic classification, this study takes advantage of both the MMTx and LVG tagging results by merging them. For example, the noun phrase "depression illness" in Figure 4-4 is transformed into a single concept in the MMTx tag, which is better than the LVG tagging results in Figure 4-5 for semantic representation purposes. On the other hand, as verbs, adjectives, and adverbs such as "get" and "better" are also indispensable for semantic analysis, their LVG tagging results are used in addition to the MMTx tags for noun phrases.

The merging of LVG and MMTx tagging results can be performed based on the beginning and ending offset of tokens because MMTx and LVG provide consistent offsets. In this study, a POS-based filter is applied to LVG tagging in order to collect only tags whose POS is noun, verb, adjective or adverb into the final tagging result. LVG tags for other POSs such as articles, prepositions, etc. are relatively less meaningful for semantic analysis purposes and hence are ignored in this study. In addition, a threshold of the MMTx mapping score is set to determine the preference of MMTx tags over LVG tags. As learned from the merge testing, a mapping score of 850 is a practical, effective threshold. When the mapping score is lower than 850, MMTx tags are likely to have a low quality and are therefore replaced by LVG tags.

In order to make the semantic tags machine readable and processable, this study defines the semantic tag syntax to include essential semantic metadata only. An example is given in Figure 4-6. The tagging result of a sentence includes three types of components: (1) the sequence number of the sentence within the text and the beginning and end offsets so that a sentence can be uniquely identified, (2) the original text of the sentence, and (3) semantic tags in either LVG format by default or MMTx format for Metathesaurus concepts. LVG tags have five fields and MMTx tags have six fields. The first field of a semantic tag is the normalized tag name. The second and the third are the beginning and ending offsets of the corresponding tokens. The fourth field records the token sequence number within the sentence. It empowers computer programs to calculate the proximity between co-occurring semantic tags in terms of the number of intervening tokens. Metadata in the fifth field varies

depending on the tag type. It saves a semantic type for MMTx tags, or POS property for LVG tags. The sixth field is used for MMTx tags only to save the MMTx mapping score.

Sentence Tag Syntax:

|===sentence===|sentenceSequenceNum|begin-offset|end-offset|#%#original sentence%#%[LVG or MMTx tag]|[LVG or MMTx tag]|. . .

LVG Tag Syntax

```
[LVG tag,begin-offset,end-offset,Term Sequence Num within the hosting
sentence,POS]
```

MMTx Tag Syntax

```
[MMTx tag,begin-offset,end-offset,Term Sequence Num within the hosting
sentence,Semantic Type,mapping score]
```

Semantic Tag Example:

```
|===Sentence===|0|0|20|#%#What is Depression?  %#%[Depressive
disorder,8,17,2,Mental or Behavioral Dysfunction,1000]
. . .
|===Sentence===|8|520|610|#%#People with a depressive illness cannot merely
"pull themselves together" and get better. %#%[Persons,520,525,0,Population
Group,1000]|[Depressive disorder,534,543,3, Mental or Behavioral
Dysfunction,1000]|[merely,560,565,6,adv]|[pull,569,572,7,adv]|[together,585,5
92,9,adv]|[get,599,601,11,verb]|[best,603,608,12,adj]
```

Figure 4-6 Final semantic tagging result

4.3 Methods of Semantic Classification

After semantic and lexical tagging, the tagged sentences are converted into semantic tag instances which are comprised of normalized semantic tags. Then in a pattern discovery phase, patterns in the semantic tag instances are obtained. According to the tag schema shown in Figure 4-6, patterns can contain different types of information including semantic components, syntactic metadata, positional relationships between semantic components, etc. During the testing phase, sentences from the testing text are likewise converted into semantic

tag instances. A learned classifier then identifies whether the semantic tag instances contain the patterns learned from the training data set so that it can classify them as TRUE or FALSE with reference to specific rating criteria.

In this study, two different systems have been developed for conducting semantic classification: a rule-based classification system and a machine learning based classification system. However, due to the data sample size and the time and funding resource limit of the thesis study, the second has been implemented as a proof of concept. The difference between them is mainly in how the classification patterns are created. Generally speaking, the rule-based classification relies on manually extracted classification patterns to classify sentences in the testing text. Human knowledge engineering has been conducted to extract from the training semantic tag instances the patterns in which different semantic units are combined together to express a rating criterion. After classification rules have been established, the system is able to classify sentences in the testing text into binary groups (i.e. TRUE or FALSE) with reference to each rating criterion, and then to further rate the quality of the testing text based on rule-based classification results. In the machine learning system, the patterns for classification are learned from training data by the machine learning algorithm, Naïve Bayes in this study. Thus, not only the rating process but also the training is automatic. The main procedures for constructing the rule-based classification system and the machine learning system are represented in Figure 4-7. Their design details are introduced in the following sections.

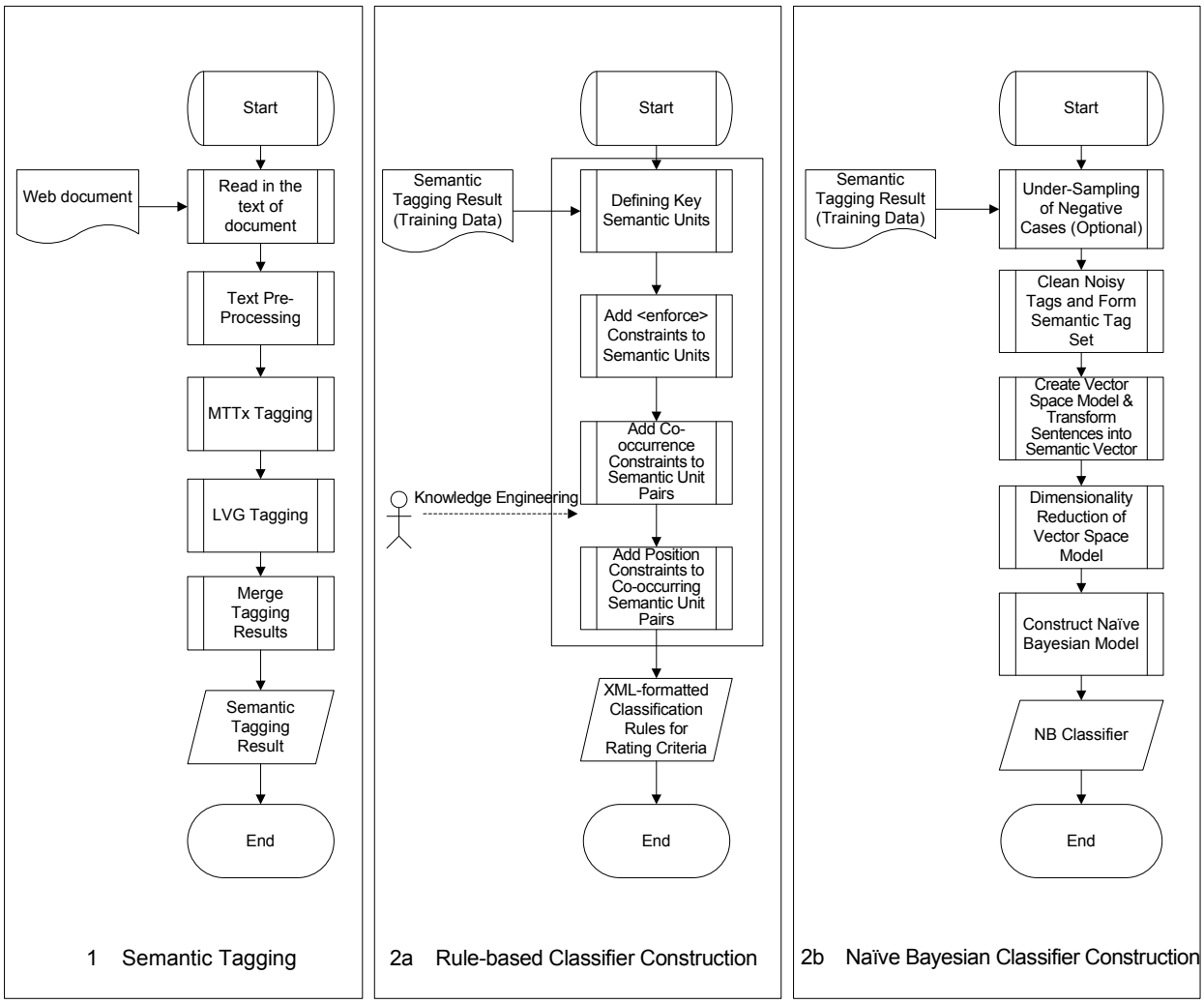

Figure 4-7 Process flow charts for semantic tagging & classification modules

4.3.1 Rule-based Classification

For each rating criterion, all of the sentences in the training data set that the human raters identified to be in concordance with the criterion are considered as positive training examples. Researchers have studied the texts of the positive examples and their semantic tag instances and have manually extracted patterns and established decision rules. During the testing phase, computer programs scan the semantic tag instances of testing sentences

looking for these patterns and employing the established rules to determine whether a sentence is in concordance with the individual rating criterion. Details of the patterns are given below.

4.3.1.1 Rule Generation

As required by the rating instructions, human raters have underlined the text components of the sentences which have led them to match a sentence with a treatment criterion. For each individual treatment criterion, such underlined components were collected and linked to the corresponding POS property and semantic entities and tags in the semantic tag instances. Token normalization and other NLP processing have unified the semantic tags for the same logic concept. Therefore, patterns established based on semantic tags are generic and can certainly fit one or multiple positive training instances even though in texts there could be different expressions in natural language. For example, the text “side effect” and “adverse effect” have the same semantic tag – “effect side”. In addition, the co-occurrence feature of semantic tags is also part of the patterns. Because the semantic tag instance includes the beginning and end position of every tag and its token sequence as well, a proximity constraint can be calculated by counting the number of tokens between two co-occurring semantic tags. The sequential relationship is defined to specify one tag occurring before and/or after the other when such a constraint exists. The proximity constraint and sequential constraint, if existing, are used to supplement the co-occurrence constraint to make the pattern more precise. While building a classification rule for a rating criterion, the maximum proximity learned from the training samples is used as the constraint threshold so that the developed patterns fit all positive training cases.

The semantic tags and patterns that are learned from the training web pages attempt to capture the resources available to authors to express the semantic content of a rating criterion. Since there can be more than one way of conveying the meaning of a depression treatment criterion, an individual rating criterion classification rule can consist of one or more patterns. The computer system relies on these patterns to classify input sentences into binary groups, i.e. TRUE or FALSE. If the computer detects at least one classifying pattern in the semantic tag instance, the sentence is classified into the TRUE group for the particular rating criterion and the sentence is considered to be semantically matching with the rating criterion. Otherwise, it is classified into the FALSE group.

4.3.1.2 Classification Rules

An XML schema has been developed for presenting classification rules so that the rules are machine readable and processable. Figure 4-8 illustrates a classification rule for rating criterion No. 6 in the XML schema.

```
<?xml version="1.0" ?>
<!--DOCTYPE RulePattern SYSTEM "D:\TestLab\MachineProcessing\MachineTesting-R11-
R12\RulePattern.dtd"-->
<RulePattern>
	<ruleID>6</ruleID>
	<patAmount>3</patAmount>

	<!-- "antidepressant", "side effect", "vary", "not"(NEGPunit), proximity(2,3)=[EITHER,5] -->
	<Pattern>
		<PID>1</PID>

		<punitAmount>4</punitAmount>

		<punit>
			<eID>1</eID>
			<keyword>Antidepressive Agents</keyword>
			<tagType>MMTx-1</tagType>
			<pos>N</pos>
			<synset>
				<synCount>3</synCount>
				<syn>
					<term>MAOIs?</term>
					<tagType>TEXT</tagType>
					<pos>N</pos>
				</syn>
				<syn>
					<term>SSRIs?</term>
					<tagType>TEXT</tagType>
					<pos>N</pos>
				</syn>
				<syn>
					<term>SNRIs?</term>
					<tagType>TEXT</tagType>
					<pos>N</pos>
				</syn>
			</synset>
			<alter_in_context>
				<altCount>2</altCount>
				<alternative>
					<term>Pharmaceutical Preparations</term>
					<tagType>Hypernym</tagType>
					<pos>N</pos>
				</alternative>
				<!-- This MMTx includes free text Medication, medicine and drug -->
				<alternative>
					<term>drug</term>
					<tagType>Hypernym</tagType>
					<pos>N</pos>
				</alternative>
			</alter_in_context>
			<enforce>1</enforce>
			<co-occurrence>
				<co-flag>N</co-flag>
			</co-occurrence>
		</punit>

		<punit>
			<eID>2</eID>
			<keyword>effect side</keyword>
			<tagType>LVG</tagType>
			<pos>N</pos>
			<synset>
```

```
                        <synCount>1</synCount>
                        <syn>
                                <term>side-?effects?</term>
                                <tagType>TEXT</tagType>
                                <pos>unknown</pos>
                        </syn>
                </synset>
                <alter_in_context>
                        <altCount>0</altCount>
                </alter_in_context>
                <enforce>1</enforce>
                <co-occurrence>
                        <co-flag>Y</co-flag>
                        <cotermContainer>N</cotermContainer>
                </co-occurrence>
        </punit>

        <punit>
                <eID>3</eID>
                <keyword>vary</keyword>
                <tagType>LVG</tagType>
                <pos>V</pos>
                <synset>
                        <synCount>2</synCount>
                        <syn>
                                <term>change</term>
                                <tagType>LVG</tagType>
                                <pos>V</pos>
                        </syn>
                        <syn>
                                <term>alter</term>
                                <tagType>LVG</tagType>
                                <pos>V</pos>
                        </syn>
                </synset>
                <alter_in_context>
                        <altCount>1</altCount>
                        <alternative>
                                <term>differ</term>
                                <tagType>LVG</tagType>
                                <pos>V</pos>
                        </alternative>
                </alter_in_context>
                <enforce>1</enforce>
                <co-occurrence>
                        <co-flag>Y</co-flag>
                        <cotermContainer>Y</cotermContainer>
                        <co-term>
                                <co-eid>2</co-eid>
                                <co-occur_proximity>5</co-occur_proximity>
                                <position_relation>EITHER</position_relation>
                        </co-term>
                        <!-- one PUNIT is allowed to have multiple co-occurring PUNITs-->
                </co-occurrence>
        </punit>

        <punit>
                <eID>4</eID>
                <keyword>not</keyword>
```

```
                <tagType>LVG</tagType>
                <pos>ADV</pos>
                <synset>
                        <synCount>3</synCount>
                        <syn>
                                <term>never</term>
                                <tagType>LVG</tagType>
                                <pos>ADV</pos>
                        </syn>
                        <syn>
                                <term>no</term>
                                <tagType>LVG</tagType>
                                <pos>ADJ</pos>
                        </syn>
                        <syn>
                                <term>no</term>
                                <tagType>TEXT</tagType>
                                <pos>unknown</pos>
                        </syn>
                </synset>
                <alter_in_context>
                        <altCount>3</altCount>
                        <alternative>
                                <term>unlikely</term>
                                <tagType>LVG</tagType>
                                <pos>N</pos>
                        </alternative>
                        <alternative>
                                <term>barely</term>
                                <tagType>LVG</tagType>
                                <pos>N</pos>
                        </alternative>
                        <alternative>
                                <term>rarely</term>
                                <tagType>LVG</tagType>
                                <pos>N</pos>
                        </alternative>
                </alter_in_context>
                <enforce>-1</enforce>
                <co-occurrence>
                        <co-flag>Y</co-flag>
                        <cotermContainer>Y</cotermContainer>
                        <co-term>
                                <co-eid>3</co-eid>
                                <co-occur_proximity>4</co-occur_proximity>
                                <position_relation>BEFORE</position_relation>
                        </co-term>
                </co-occurrence>
        </punit>
    </Pattern>

    <Pattern>
        . . .
    </Pattern>

</RulePattern>
```

Figure 4-8 Classification rule for rating criterion #6

As shown in Figure 4-8, the metadata set of a classification rule is included in a tag pair called <RulePattern>. Metadata <ruleID> indicates which rating criterion the rule is for. The <patAmount> shows the number of patterns extracted from the training data. A full description of a classifying pattern is saved in the <Pattern> tag pair. <Pattern> has self-explanatory tags. <PID> is the pattern identifier and <punitAmount> indicates how many semantic units (i.e. <PUNIT>s) comprise the pattern.

A <PUNIT> defines a semantic unit in a pattern. During the pattern matching process, computer programs sequentially scan the semantic tags of a sentence to check if the specified PUNIT occurs in the sentence and whether the constraints of the PUNIT are complied with by the sentence. A PUNIT is defined using the following metadata:

- <eID> --- the sequence number of the PUNIT in the current pattern
- <keyword> --- the normalized term of the PUNIT, i.e. the semantic tag
- <tagType> --- the tag type of PUNIT, valid values include MMTx, LVG tag, Hypernym, and TEXT;
  TEXT is a supplement to MMTx and LVG tags to handle the cases when the semantic tagging occasionally skips over a token or phrase. When the tagType is TEXT, the keyword field uses a regular expression to provide a flexible means to match strings of text.
  Hypernym is the hypernym of the keyword.
- <POS> --- the part of speech of the PUNIT in the sentence; it could have multiple choices. For example, 'V|N' means either verb or noun.
- <synset> --- a set of synonyms of the <keyword>.
- <alter_in_context> --- a set of alternative expressions which are considered as synonyms of PUNIT only if used in the context of depression treatment.

- <enforce> --- the confidence of this PUNIT; it could be -1, or in the range (0, 1]. If enforce=1, then this PUNIT must be found in a sentence to match the pattern. If 0 < enforce < 1, then the pattern can still possibly be matched when this PUNIT is absent as long as other mandatory PUNITs are satisfied by the sentence. In such a case, the confidence of the pattern matching will be an iterative product of the confidence and the enforce value of each PUNIT. Confidence has an initial value of 1. In addition, enforce could be -1 to define a PUNIT as a NEGATOR of the pattern. That is, if the PUNIT is found in a sentence, then the pattern is certainly not a match. It is quite often that a positive proposition and a negative proposition have common semantic tags, while the negative proposition has an extra "not" or other negation expression.

  The <enforce> value for each PUNIT has been obtained based on the frequency of the PUNIT occurrence in training cases under the same pattern type. For example, among 10 training instances which do not have a certain PUNIT while having other PUNITs satisfied, if 7 instances are positive, then enforce = 0.7.

- <co-occurrence> --- indicates the co-occurrence of this PUNIT with another PUNIT
- <co-flag> --- a flag to indicate the co-occurrence, either Y (yes) or N (no)
- <co-term> --- the co-occurring PUNIT
- <co-eid> --- the eID of the co-occurring PUNIT
- <co-occur_proximity> --- specifies the distance between two co-occurring PUNITs. The distance is the number of tokens in between the co-occurring PUNITs.
- <position_relation> --- the sequential relationship of co-occurring PUNITs. The current PUNIT could be BEFORE, AFTER, or BOTH the co-occurring PUNIT.

In terms of pattern development, a pattern is comprised of a minimum set of PUNITs which are necessary conditions for identifying a rating criterion. In a sentence, some semantic components represent relatively subtle aspects of the sentence meaning. Such components, for example adverbs for degree such as "some", "basically", etc., are not included as PUNITs

in a classification pattern, if the removal of these components does not change the determination of classification type of this sentence. The purpose is to make the generated pattern be generic enough and fit as many positive training sentences as possible. During the matching process, if a sentence contains all PUNITs in a pattern and satisfies all constraints, this sentence is considered to be an exact match. In order to increase recall, not only sentences with “exact match” but also those having a high degree of similarity with the learned patterns are classified to be TRUE with reference to a specific rating criterion. The matching confidence value is used to represent the similarity between a testing sentence and a learned pattern. In the XML formatted pattern definition file, each PUNIT has been assigned an enforce rate. The matching confidence of a whole sentence is the product of the enforce rate of all PUNITs. The more PUNITs matched by the semantic tag instance of a sentence, the higher the degree of confidence. In this study, the threshold of matching confidence has been set to be 0.75. It has been determined that any sentence which has a matching confidence higher than 0.75 is considered to be in concordance with a rating criterion. On the other hand, there could also be certain cases in which the meaning of a sentence is not consistent with the rating criteria although the sentences contain the PUNITs required. The constraints specification in the pattern, such as POS, co-occurrence, proximity, negation, and etc., has been designed for filtering out false positive cases as much as possible. This has allowed achieving reasonable recall without losing precision.

4.3.1.3 Rule Matching

A JAVA program has been implemented to read both the semantic tag instances and the classification rules. For each sentence, the patterns of a rule are checked in sequence. If any

pattern is matched, the sentence is classified into the TRUE group. If no pattern can be matched, the sentence is classified into the FALSE group. During the process of pattern matching, PUNITs are validated in sequence. Pattern matching stops if any of the following three conditions are met: pattern matching confidence factor is lower than the threshold, a MANDATORY PUNIT (i.e. enforce=1) is not found in the sentence, or a NEGATIVE PUNIT (i.e. enforce=-1) is found in the sentence. Examples in Figure 4-9 illustrate a successful and a false positive matching case respectively. A detailed discussion of the rule-based classification results and a performance evaluation can be found in Chapter 5.

```
Criterion #6 – "The side effect profile varies for different antidepressants."

Examples of sentences predicted as TRUE with reference to criterion #6:

Case 1 – successful matching case
They generally have more side effects than newer (second-generation)
antidepressants such as serotonin reuptake inhibitors (SSRIs) and other
second-generation antidepressants such as bupropion (Wellbutrin, Wellbutrin
SR) and duloxetine (Cymbalta).

Semantic Tag Instance:
|===Sentence===|1|207|455|#%#They generally have more side effects than newer
(second-generation) antidepressants such as serotonin reuptake inhibitors
(SSRIs) and other second-generation antidepressants such as bupropion
(Wellbutrin, Wellbutrin SR) and duloxetine (Cymbalta).
%#%[generally,212,220,1,adv]|[more,227,230,3,adv]|[effect
side,232,243,4,noun]|[new,250,254,6,adj]|[generation
second,257,273,7,adj]|[Antidepressive Agents,276,290,8,Pharmacologic
Substance,1000]|[Serotonin Uptake Inhibitors,300,308,10,Pharmacologic
Substance,1000]|[reuptake,310,317,11,noun]|[inhibiter,319,328,12,noun]|[antid
epressant,331,335,13,noun]|[different,342,346,15,adj]|[generation
second,348,364,16,adj]|[antidepressant,366,380,17,noun]|[Bupropion,390,398,19
,Organic Chemical, Pharmacologic
Substance,1000]|[antidepressant,401,410,20,noun]|[antidepressant,413,422,21,n
oun]|[sr,424,425,22,noun]|[duloxetine,432,441,24,Organic Chemical,
Pharmacologic Substance,1000]

Case 2 – false positive matching case
Side effects of tricyclics, which vary from person to person, may include dry
mouth, blurred vision, constipation, problems passing urine, sweating, light-
headedness and excessive drowsiness.

Semantic Tag Instance:
|===Sentence===|3|365|557|#%#Side effects of tricyclics, which vary from
person to person, may include dry mouth, blurred vision, constipation,
problems passing urine, sweating, light-headedness and excessive drowsiness.
%#%[effect side,365,376,0,noun]|[Antidepressive Agents,381,390,2,Organic
Chemical, Pharmacologic
Substance,1000]|[vary,399,402,4,verb]|[Persons,409,414,6,Population
Group,1000]|[Persons,419,424,8,Population
Group,1000]|[include,431,437,10,verb]|[dry
mouth,439,447,11,noun]|[Vision,458,463,12,Organism
Function,861]|[Constipation,466,477,13,Sign or
Symptom,1000]|[disturbance,480,487,14,noun]|[pa,489,495,15,verb]|[urine,497,5
01,16,noun]|[sweat,504,511,17,verb]|[headedness
light,514,529,18,noun]|[excessive,535,543,20,adj]|[Drowsiness,545,554,21,Sign
or Symptom,861]
```

Figure 4-9 Matching result according to classification rule

4.3.2 Machine Learning – Naïve Bayes Classification

The rule-based classification has required human effort to extract patterns and to establish classification rules. In addition to this approach, a machine learning based method (Naïve

Bayes) has also been designed in this study. Compared to the approach using rule-based classification, the implementation of Naïve Bayes classification empowers a computer system to learn patterns and to train the classifier, without the need for human effort to generalize the classification patterns.

By presenting the rule-based quality rating, this study tries to demonstrate that a computer system, by utilizing shallow semantic processing and analysis, can rate health care information quality directly based on shallow text semantics. The purpose of developing a machine learning based classification system is to further demonstrate that semantic processing and the tagging results can be integrated into fully automated algorithms. Thus, not only can quality rating be done automatically, but the training process can also be automated.

Due to the time and resource limitations of this thesis study, the sample size of depression treatment web pages is small: 201. This choice of sample size has kept the amount of human rating work reasonable. Regarding the matching of rating criteria by the human raters in the corpus of 201 web pages, some treatment criteria had few or no matches (positive cases). Since the machine learning technique requires a reasonably large training data set to contain a sufficient number of positive cases (TRUE instances in this study), the machine learning based classification has been tested with a subset of rating criteria, rather than the full set. For this reason, the implementation and testing of the machine learning based rating has been conducted on a reasonably small scale to provide proof of concept for full automation – three depression treatment criteria, Nos. 1, 6, and 12-B, have been selected from the whole set for

demonstration. These three rating criteria have been selected for two reasons: they represent different semantic complexity, and the selected treatment criteria have relative large population of human-identified positive sentences in order to have enough positive instances for training classifiers using machine learning algorithm. The first reason is an attempt to reduce any bias on the performance of the testing phase caused by the semantic complexity of the rating criteria.

For quality rating purposes, the computer program has to identify references to the three depression treatment criteria (i.e. #1, #6, #12-B) that may be present on each web page, and compute the quality score, which represents the number of rating criteria addressed in the web page. Thus, the rating score ranges from 0 to 3. The procedure can be divided into three steps. First, machine learning is performed for each individual rating criterion in order to build a classifier dedicated to that rating criterion. The training data are the semantic tag instance of human-rated sentences from the training web pages. Second, the resulting dedicated classifiers are applied to the testing data. Each semantic tag instance that results from the testing web page sentences is classified by the dedicated classifier as either TRUE or FALSE. Finally, after sentence classification is completed, a computer program identifies which of the three rating criteria are referred to by each web page. The number of rating criteria referred to constitutes the quality score for the machine learning approach.

4.3.2.1 Supervised Learning – The Naïve Bayes Classifier

In this study, via the machine learning approach, the classifier can be modeled as a function of the form f: X -> Y, in which Y is a Boolean-valued random variable, i.e. TRUE or

FALSE with reference to a rating criterion, and X is an array of attributes of a sentence instance. A Naïve Bayes classifier is developed in this study because it requires a small amount of training data to estimate the parameters necessary for classification and also because this method is simple to implement. A Naïve Bayes classifier is a probabilistic model that assumes that all attributes of the examples are independent of each other given the context of the class. Because of the independence assumption, the parameters for each attribute can be learned separately, and this greatly simplifies learning, especially when there is a large number of attributes. The Naïve Bayes model has many variations. In this study, the adopted model is the multi-variate Bernoulli event model, which has been used for text classification in numerous studies (McCallam & Nigam, 1998; Billsus & Pazzani, 1999; Schneider, 2003; Chen et. al., 2009).

Formula (1) is the mixture model for predicting the sentence class. The mixture model consists of mixture components c_j, $j \in \{1, \ldots |C|\}$. In our case, $|C| = 2$ because of the one-to-one correspondence between classes and mixture model components. Each component is parameterized by a disjoint subset of θ. Thus a sentence, s_i, has probability of class C_j as:

$$P(s_i|\theta) = \sum_{j=1,2} P(c_j|\theta)P(s_i|c_j;\theta) \quad (1)$$
$$= P(c_1|\theta)P(s_i|c_1;\theta) + P(c_2|\theta)P(s_i|c_2;\theta)$$

$S = \{s_1, \ldots s_{|S|}\}$ represents the training set of semantically tagged sentences which was used to train the parameters of the classification model. The class prior parameters, θc_j, can be obtained by:

$$\theta c_j = P(c_j \mid \theta) = \sum_{i=1}^{|S|} P(c_j \mid S_i) / |S| \quad (2)$$

Within the multi-variate Bernoulli model, the semantic tag instance of a sentence can be interpreted as a vector of binary attributes that indicates the occurrence or non-occurrence of a set of pre-defined semantic tags. The number of times a tag occurs in a sentence and the order of the tags are not captured. In this context, T stands for the semantic tag vocabulary established based on training sentences. Subscript *m* represents the dimension size of T, and a sentence vector is comprised of *m* different semantic tags, i.e. $\{T_1, T_2, \ldots T_m\}$. The probability of a sentence is the result of multiplying the probability of all attributes, including the probability of both occurring and non-occurring tags in the sentence.

$$P(s_i|c_j;\theta) = \prod_{k=1}^{m} (P(s_i \mid T_k)\, P(T_k|C_j;\theta)) \qquad (3)$$

Since the binary class group (i.e. either TRUE or FALSE) was assigned to sentences, $P(s_i \mid T_k)$ can only be either 0 or 1. Thus formula (3) can be transformed to:

$$P(s_i|c_j;\theta) = \prod_{k=1}^{m} (B_{ik}P(T_k|C_j;\theta) + (1-B_{ik})(1-P(T_k|C_j;\theta)) \qquad (4)$$

4.3.2.2 Classification Tool

There are many open source machine learning tools available. WEKA (Witten et al., 2011) has been chosen for this study. WEKA is a comprehensive machine learning toolkit developed at the University of Waikato in New Zealand. It implements many machine learning approaches in the JAVA programming language and it has been widely and successfully used in other machine learning related research (WEKA, 2011). Additionally, WEKA has easy-to-use JAVA APIs.

In WEKA, a data set is a collection of sample data. Each data item is called an instance. Each instance consists of a number of attributes, any of which can be nominal, numeric or a string. The external representation of a data set is an ARFF file, which consists of a header describing the attribute types and the instances as comma-separated lists (WEKA, 2011). In the context of this study, each semantic tag instance is transformed into a WEKA instance. Both training and testing data are saved in ARFF files as external input. Figure 4-10 illustrates a snippet of an ARFF file of training data for rating criterion #6. The generation of an ARFF file involves multiple processes, including constructing a vector space based on semantic tag instances, cleaning noisy semantic tags, reducing the number of dimensions, projecting the semantic tag instances to the reduced vector space, etc. Such processes are introduced in Sections 4.3.2.3 and 4.3.2.4.

```
% 1. Title: R6 Training Data
%
% 2. Source:
%    (a) Creator: Yanjun Zhang
%    (b) Date: Jan, 2011
%
@RELATION Rule6-Classification

@ATTRIBUTE treat numeric
@ATTRIBUTE major numeric
@ATTRIBUTE tag_class numeric
@ATTRIBUTE Antidepressive_Agents numeric
@ATTRIBUTE Pharmaceutical_Preparations numeric
@ATTRIBUTE atypical numeric
@ATTRIBUTE affect numeric
. . .
@ATTRIBUTE exam numeric
@ATTRIBUTE couple numeric
@ATTRIBUTE visit numeric
@ATTRIBUTE successfully numeric
@ATTRIBUTE class {Y, N}

@DATA
1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,N
. . .
0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,
0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,Y
. . .
```

Figure 4-10 ARFF file example: training data for rating criterion #6

4.3.2.3 Vector Space Model

The attributes and instances in the ARFF file can be mapped to a vector space. The attributes correspond to the dimensions of the vector space and each ARFF instance corresponds to a sentence vector. The construction of a sentence vector is based on the semantic tag instance (e.g. Fig 4-4) of this sentence. Sentences across all training web pages are collected together to form the vector space. Each sentence is represented as an independent vector instance. The number of dimensions of the vector space is determined by the number of unique semantic tags in the training data set. In this study, data cleaning is completed to remove noisy tags before refining dimensions. Two types of noisy tags are removed, as shown in Table 4-1. One has POS in numeric. These are removed because they are not useful in the classification of the three criteria in question. The other type is the tagged label containing "www". This type of tag corresponds to URL in the text content and is irrelevant information regarding depression treatment. After cleaning noisy tags, a vector space is built based on the unique semantic tags.

Table 4-1 Examples of noisy tag to be removed

Tag Content	**Tag Result / POS**
4ppd 800 944	numeric (POS)
805	numeric (POS)
10^{th}	numeric (POS)
`a actionset ca healthlinkbc htm kbase tb1939 www`	`a actionset ca healthlinkbc htm kbase tb1939 www`

In the next step, an additional dimension, i.e. the classification type of sentence instances, is added into the vector space. The classification type is either TRUE or FALSE, hence the dimension value is correspondingly Y or N depending on the human raters' labeling results.

This dimension also corresponds to the last attribute, i.e. “class” in the ARFF file. After this, the vector space is considered as the full-size dimension vector space.

4.3.2.4 Dimension Reduction

After the above processing, the vector model is supposed to be ready for computation. However, due to the huge training data size and the content scope of the web text samples, the vector space dimension size can be too large and hence the vector space can become very large and sparse. The sparse vector space is not desirable because it cannot only cause computational inefficiencies, but also impact classification performance since the weights of the feature dimensions are overwhelmed by the sparse dimensions (Kim et. al., 2005). Thus, dimension reduction has been implemented in this study.

In the following, rating criterion #1 is used as an example for illustrating the dimension reduction. Without reduction, the vector space has a dimension size of 3963 based on both positive and negative training cases. However, if only positive cases of rating criterion #1 are used exclusively to construct a vector space, the number of unique semantic tags is only 318. The reason for the difference of dimension size is straightforward—in addition to discussing a depression treatment criterion, an article can also contain some background information and knowledge about depression in sub-topics other than treatment, such as depression research groups and activities, methods for depression self-diagnosis, health care resources, etc. Such information is not semantically relevant to the specific rating criterion in question, but still contributes additional semantic tag dimensions to the full-size dimension vector space. Since some of these additional semantic tags may not be needed for representing the

rating criteria, the number of vector space dimensions can be safely reduced by removing those corresponding to irrelevant semantic tags. This study uses cosine similarity between a sentence vector and a representative of the positive case vectors to identify such removable dimensions.

In traditional text document classification studies (Billsus & Pazzani, 1999; Kim et al., 2006), cosine similarity indicates how close one vector instance is to another, with both vectors representing document instances. In this study, cosine similarity of the vectors standing for the semantic tag instances is used to represent how close one sentence is to the known positive training sentences. The more semantic components that a sentence and a positive case commonly have, the higher the similarity is.

In most traditional text classification studies, which are mainly at the document level (e.g. Han & Karypis, 2000; Guan. et al. 2009), the centroid of the positive cases is used in the similarity calculation. In this study, the objects being classified are semantic tag instances. In the vector space model, a semantic tag instance is represented by a vector instance; a semantic unit in the sentence is represented by the vector instance's value in a specific dimension. Since there can be more than one way to convey the same meaning in writing, different positive training sentences for a single criterion can possibly represent different patterns of assembling semantic units. Consequently, a vector instance of positive case A can be much different than the instance of positive case B and hence they have low similarity with each other. In this context, it would be senseless to use the centroid of the positive training sentences to calculate similarity since the features embedded in different positive

cases would offset each other. Therefore in this study, a given semantic tag instance, say V_i, is compared to each positive training instance P_j to calculate cosine similarity. The maximum value of similarity across all positive training instances is assigned to the current instance. It represents how close this instance is to the most similar positive training case. Given any sentence, its similarity ranges from 0 to 1. Assuming the positive set P has N instance cases, the similarity formula is:

$$\begin{aligned} SIM(V, P) &= \max(Sim(V, P_i)) \qquad i \in (1, 2, \ldots N) \\ &= \max((V \cdot P_i) / |V||P_i|) \end{aligned} \tag{5}$$

Based on the similarity calculated above, the following method has produced a favourable reduction of the dimension size: Using a cosine similarity measure threshold, training instances are divided into a high-similarity and a low-similarity group. If a semantic tag has the value 0 across all training instances in the high-similarity group, this indicates that the semantic tag does not occur in any training sentences which have semantic components similar to positive cases. These all-0's dimensions are removed from the model and training instance vectors are likewise modified. Indeed, a manual review of these removed dimensions has confirmed that their corresponding semantic concepts are semantically unrelated to depression treatment. Some typical examples of these concepts are cancer, clinic, university, etc. In this study, different similarity thresholds have been tested for reducing vector space dimensions, ranging from 0.4 to 0.8 in intervals of 0.05. The system has satisfactory performance when the similarity threshold is 0.5. For depression treatment criterion #1, the dimension size was reduced to 635 from the original 3963. After dimension reduction, the obtained vector space is the final model to be used. Dimensions in this final

vector space model correspond to the attributes in the ARFF files. Each rating criterion has its own vector space model, and the training and testing ARFF files can be generated by respectively projecting the semantic tag instances onto the specific vector space.

4.3.2.5 Sentence Classification Algorithm

The algorithm for the semantic classification follows.

1. For each sentence from all of the training web pages create a semantic tag instance.
2. Accumulate all of the semantic tags from all of the instances created in step 1 into a set called *T*. Clean *T* by removing the pre-defined noisy tags (see Section 4.3.2.3).
3. Identify the unique tags in *T* and define a vector space: each dimension represents a unique tag in *T*.
4. For each rating criterion, *t*, do steps a to d.
   a. Given training instances, create an ARFF file (see Figure 4-10) for *t* by doing steps i to vi.
      i. Create a vector instance for each semantic tag instance: For each dimension in the vector, a value 1 means that the semantic tag occurs in the semantic tag instance, a value 0 means that it does not. (A simplified example is shown in Figure 4-11.)
      ii. Add an extra dimension at the end of each vector instance to represent the human rater classification type of the sentence for the treatment criterion t. A value Y indicates the classification is TRUE; a value N, FALSE.
      iii. Let the set POSITIVE be all vector instances with classification type TRUE.
      iv. For every vector instance, find the maximum of the similarity between this instance and each member in POSITIVE.
      v. Using the similarity threshold to identify irrelevant dimensions and remove them from the vector space model (see Section 4.3.2.4).
      vi. Project vector instances from the initial vector space to the reduced space.
   b. Use the training data set (the ARFF file from step a) to train the classifier.
   c. For each testing data sentence, *s*, do steps i to ii.
      i. Create a testing vector instance, v, using the vector space definition generated in step 4.a.v.
      ii. Determine the classification type for v using the trained classifier from step b.
   d. For all testing data sentences, compare the predicted classification type to the human rater label; calculate the prediction accuracy statistics.

A semantic tag instance:

```
[generally], [more], [effect side], [new], [generation second],
[Antidepressive Agents]
```

Vector space definition:

```
[new], [generation second], [Antidepressive Agents], [generally], [more],
[effect side], [old], [depression], [sleep], . . .
```

Vector presentation of the semantic tag instance:

```
[1, 1, 1, 1, 1, 1, 0, 0, 0, . . .]
```

Figure 4-11 A simplified example of a semantic tag instance and its vector representation

4.4 Quality Score of Web Pages

The quality scoring and rating procedure is implemented in a JAVA program. For every rating criterion, the learned classifiers (both the rule-based and machine learning based classifier) predict each testing sentence to be either TRUE or FALSE. After going through all sentences on a testing web page, the computer program knows whether a treatment criterion is contained in that web page. If the answer is yes, the computer program assigned a score of 1 to the web text. By accumulating the scores of all three rating criteria, the machine obtains the information quality score of a web page. The logic for calculating quality score is the same procedure that the human rater used to assign the quality score. Through statistical analysis, the quality rating results from the computer are comparable to the human raters'. The details of the statistical analysis are provided in the next chapter.

Chapter 5
Performance Evaluation and Data Analysis

Chapter 5 reports the performance evaluation of the quality rating approach. The quality rating results produced by both the rule-based and machine learning approaches are presented. The quality rating effectiveness of these approaches is examined using quantitative analysis. In addition, case analysis of both success and failure is conducted to illustrate how well the semantics-based rating approach works and what types of challenges confront this approach.

5.1 Evaluation Approach

The testing data set in this study comprises 31 web pages. As introduced in section 3.6, stratified random sampling was used to select the test web pages. Both of the automatic quality rating approaches (i.e. rule-based approach and machine learning approach) were applied independently to rate the content quality of every test page.

Table 5-1 Frequency distribution of web pages

Number of Rating Criteria in a Web Page	Collected Web Pages	Testing Pages
7 – 8	18	3
5 – 6	26	4
3 – 4	51	8
1 – 2	83	13
0	23	3
Total	201	31

The rule-based approach was applied to all 23 rating criteria; while, as explained in Section 4.3, the testing of the machine-learning approach included criteria 1, 6, and 12-B only. As

defined in the previous chapter, a quality score of a web page represents the number of unique rating criteria contained in that page. Thus, for the rule-based approach the quality scores could range from 0 to 23, while for the machine learning approach the maximum possible quality score was 3 (the number of criteria examined). The quality scores of each testing page for the two approaches are shown in Table 5-2 and Table 5-3.

Table 5-2 Quality score assigned to testing web pages by rule-based approach

Testing Page ID	Quality Score via Human Rating	Quality Score via Rule-Based Rating	Quality Score Difference *
1	7	7	0
2	7	6	-1
3	8	7	-1
4	6	5	-1
5	6	6	0
6	5	5	0
7	5	4	-1
8	4	5	1
9	3	4	1
10	4	3	-1
11	3	4	1
12	3	4	1
13	4	4	0
14	3	2	-1
15	2	5	3
16	2	3	1
17	2	2	0
18	2	2	0
19	2	2	0
20	3	2	-1
21	2	2	0
22	2	1	-1
23	2	1	-1
24	1	2	1
25	1	1	0
26	1	1	0
27	1	1	0
28	1	0	-1
29	0	0	0
30	0	0	0
31	0	0	0
Total	93	91	Not Applicable

Note:
The quality score was assigned based on all the rating criteria.
* Quality score difference = quality score via rule-based rating - quality score via human rating

Table 5-3 Quality score assigned to testing web pages for criteria #1, #6, and #12-B

Testing Page ID	Quality Score via Human Rating	Quality Score via Machine Learning Rating	Quality Score Difference (machine learning vs. human rating)*	Rule-Based Rating Result	Quality Score Difference (rule-based rating vs. human rating)**
1	2	3	1	2	0
2	3	3	0	3	0
3	3	3	0	3	0
4	3	3	0	3	0
5	2	2	0	2	0
6	3	3	0	2	-1
7	3	3	0	3	0
8	2	2	0	2	0
9	2	2	0	2	0
10	3	3	0	2	-1
11	1	1	0	1	0
12	1	1	0	2	1
13	2	3	1	2	0
14	1	1	0	0	-1
15	1	1	0	1	0
16	1	1	0	1	0
17	1	1	0	1	0
18	1	1	0	1	0
19	1	2	1	1	0
20	1	2	1	1	0
21	2	3	1	2	0
22	1	2	1	1	0
23	1	1	0	1	0
24	0	1	1	1	1
25	1	2	1	1	0
26	1	1	0	1	0
27	1	1	0	0	-1
28	0	2	2	0	0
29	0	0	0	0	0
30	0	1	1	0	0
31	0	0	0	0	0
Total	44	55	Not Applicable	42	Not Applicable

Note:
The quality score was assigned based on criteria #1, #6, and #12-B only.
* The quality score difference = quality score via machine learning - quality score via human rating.
** The quality score difference = quality score via rule-based rating - quality score via human rating.

5.2 Page Quality Score Results

According to the human rating results, the page with the highest quality score contains 8 different rating criteria, while the page with the lowest quality score includes none of the rating criteria. The human raters identified a total of 92 criteria across the testing web pages.

The quality scores generated by the rule-based approach ranged from 0 to 7, and a total of 91 criteria were identified across the testing web pages using this approach. Table 5-2 shows the quality scores rated by rule-based system. For 14 of the 31 pages (45.2%) the rule-based scores and the human rating quality scores were identical. In 10 pages (32.3%) the rule-based quality scores were one lower than the human rating quality scores, and in another 6 out of 31 pages (19.4%) the rule-based quality scores were higher by one. For only one page (3.2%) was the difference between the rule-based and human scores greater than one (testing page no. 15, rule-based score higher by 3).

The rule-based rating results were very close to human rating results not only in terms of quality score (i.e. the total number of unique rating criteria identified in each web page), but also in the specific criteria identified in the pages (see Figure 5-1). The large majority of the criteria identified by human raters were also identified using the rule-based approach (83.7% of true criteria, or 77 out of 92), and only 16.3% (i.e. 15 out of 92) of the criteria identified by the human raters were missed by rule-based approach. Among the 91 criteria identified by the rule-based approach, 13 (or 14.3%) were ‘false positives’ in that they were not identified by or accepted by human raters.

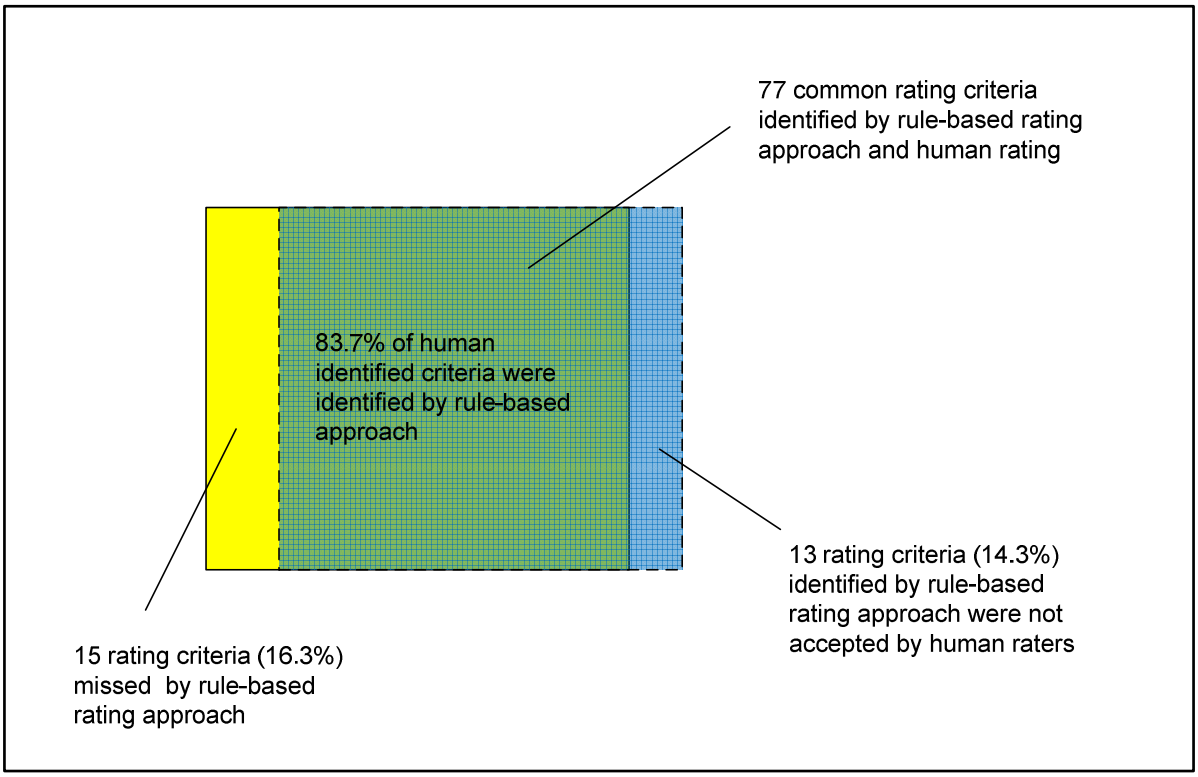

Figure 5-1 Identified rating criteria (rule-based rating vs. human rating)

The correlation between the rule-based quality scores and the human rating quality scores is shown in Figure 5-2. Pearson correlation between these two sets of quality scores is used to evaluate the performance of the rule-based machine rating approach in comparison to evidence-based human rating results. The linear correlation between two measures is positive, strong, and statistically significant ($r = 0.909$, or $r^2 = 0.827$, $p < .001$). $r^2 = 0.827$ means that 82.7% of the variance of the quality scores generated by rule-based approach is associated with the variance in the quality scores generated by human raters. In Griffiths et al. (2005), which used keyword-based automatic approach to evaluate the quality of thirty websites, the correlation between the automatically rated quality scores and the evidence-based human ratings was $r=0.850$ ($r^2 = 0.723$, $p < .001$).

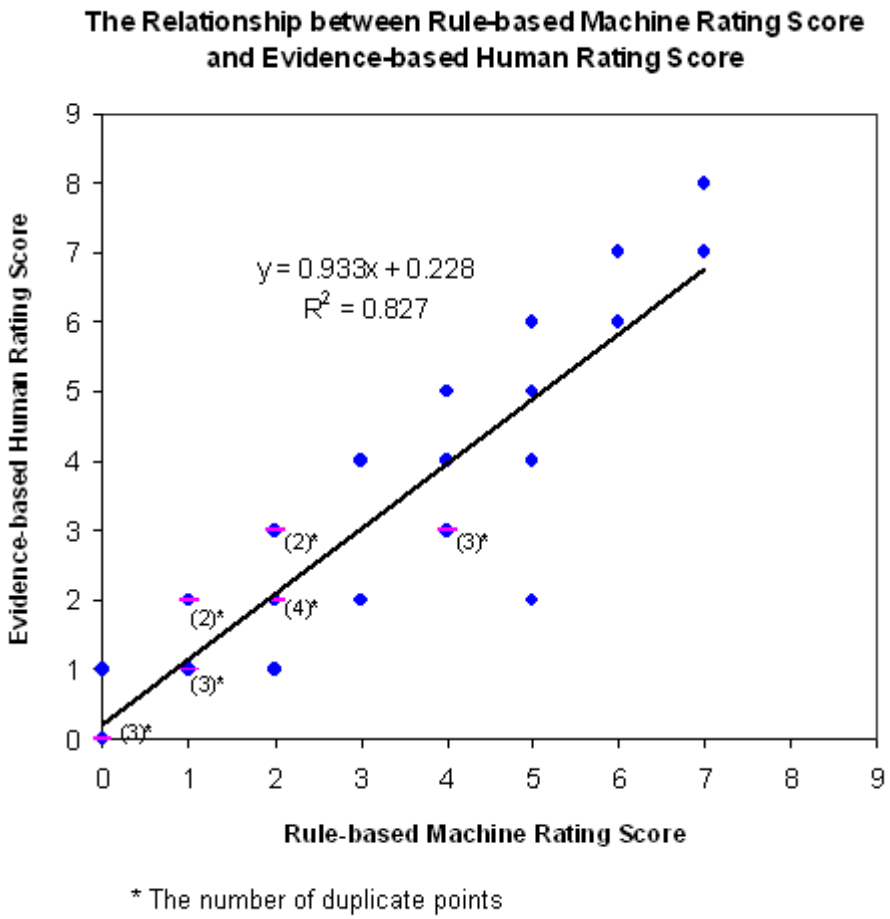

Figure 5-2 Relationship between rule-based quality scores and human rating quality scores

The quality scores generated by the machine learning approach using criteria #1, #6 and #12-B are listed in Table 5-3. The quality scores generated by the machine learning (i.e. Naïve Bayes) approach ranged from 0 to 3, and a total of 55 criteria were identified across the testing web pages using this approach. For 21 of the 31 pages (67.7%) the machine learning quality scores and the human rating quality scores were identical. In 9 pages (29.0%) the machine learning quality scores were one higher than the human rating quality scores. For only one page (3.2%) was the difference between the machine learning and human rating quality scores greater than one (testing page no. 28, machine learning quality score higher by

2). The same statistics contrasting human rating to rule-based rating results on criteria #1, #6 and #12-B were also conducted. For 25 of the 31 pages (80.6%) the rule-based scores and human rating quality scores were identical. In 4 pages (12.9%) the rule-based scores were one lower than the human rated quality scores, and in another 2 out of 31 pages (6.5%) the rule-based scores were one higher than the human rated quality scores.

The accuracy of criteria identification using the machine learning approach is shown in Figure 5-3. All the criteria identified by human raters were successfully identified using the machine learning approach. The machine learning approach, however, identified 11 extra criteria being false positive. Given the same set of criteria and web pages, rule-based rating identified 90.9% (i.e. 40/44) of criteria identified by human raters, with 4 (i.e. 9.1%) human identified criteria being missed. Among the 42 criteria identified by the rule-based approach, 2 (i.e. 4.8%) were 'false positive' in that they were not accepted by human raters.

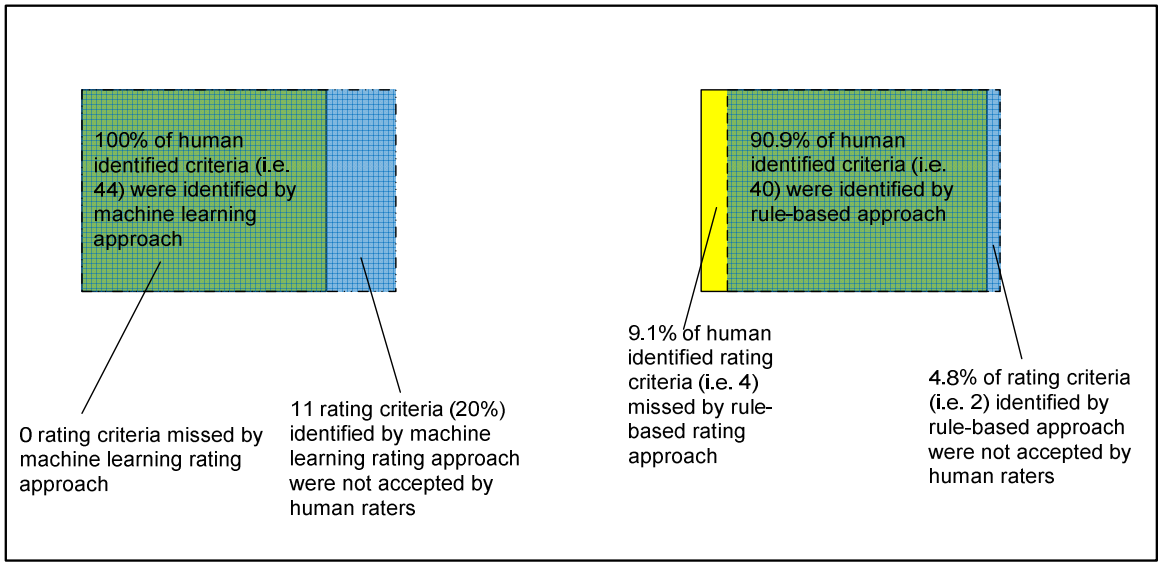

Figure 5-3 Identified rating criteria (#1, #6, and #12-B)

The linear correlation between machine learning based quality scores and the evidence-based human rating quality scores was high and statistically significant ($r = 0.841$, $r^2 = 0.707$, $p < .001$, see Figure 5-4). The high linear correlation results suggest that either automated approach may be used for evaluating the quality of online health care information.

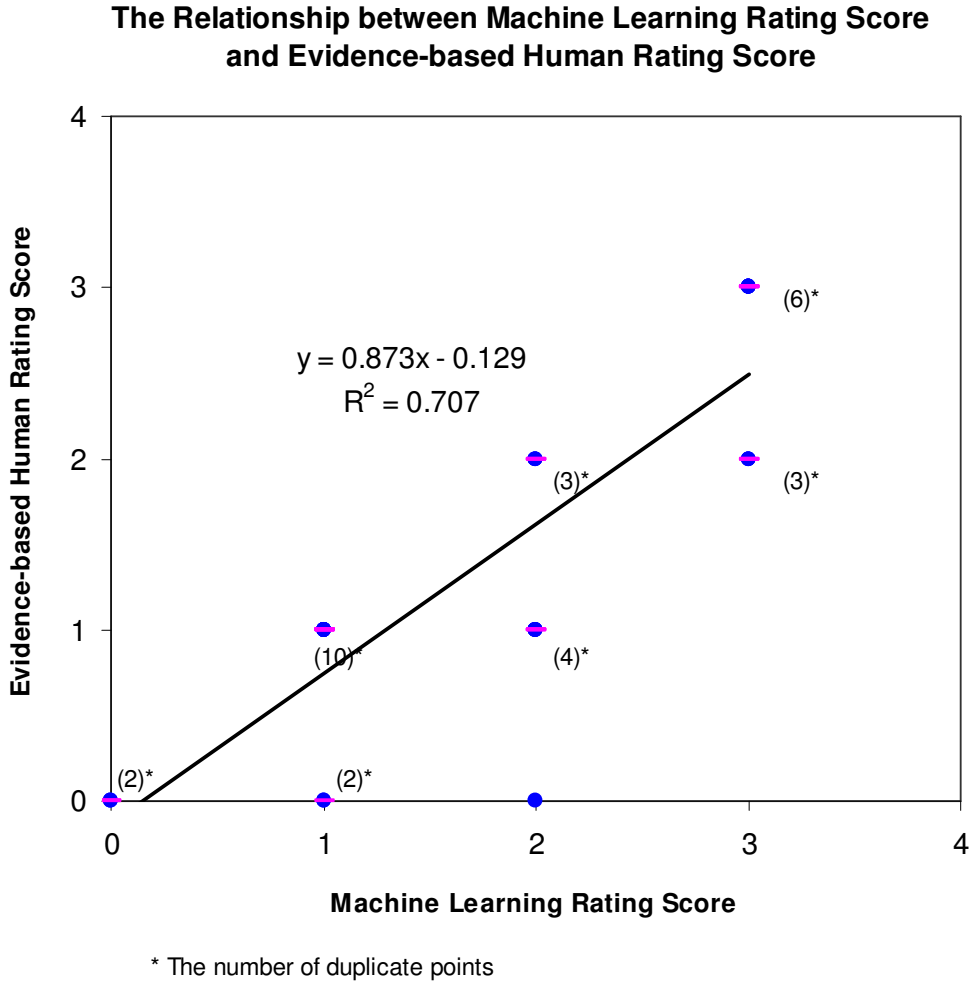

Figure 5-4 Relationship between machine learning quality rating scores and human rating quality scores

The following section illustrates the cases in which computer programs performed well, along with the false negative and false positive cases in which the computer program results did not match the human assessments.

5.3 Cases Analysis

As introduced in Chapter 3 and 4, the computer rating programs process the web pages sentence by sentence. The computer programs read a single sentence and classify it relative to each rating criterion, identifying whether the sentence is an instance of each criterion. The matching is a binary classification process based on recognizing and processing semantic components inside the sentence. The example cases used for the analysis in this section are actual sentences in plain text format pulled from the testing web pages. These examples were identified by human rater and/or computer programs as matching one of the rating criteria.

5.3.1 Successful Cases

The cases presented in Figure 5-5 are examples of the sentences that computer systems identified as criteria #1, #6, #20 in different testing pages.

Successful cases:

Rating criterion #1:

"Antidepressant medication is an effective treatment for major depressive disorder."

Testing page PID=25

1. SSRIs affect mainly serotonin and have been found to be effective in treating depression and anxiety without as many side effects as some older antidepressants.

Rating criterion #6:

"The side effect profile varies for different antidepressants."

Testing page PID=1

2. SSRIs and SNRIs are more popular than the older classes of antidepressants, such as tricyclics-named for their chemical structure-and monoamine oxidase inhibitors (MAOIs) because they tend to have fewer side effects.

Testing page PID=2

3. Side effects may vary depending on the medicine you take, but common ones include stomach upset, loss of appetite, diarrhea, feeling anxious or on edge, sleep problems, drowsiness, loss of sexual desire, and headaches.

Testing page PID=4

4. However, because TCAs tend to have more numerous and more severe side effects, they're often not used until you've tried SSRIs first without an improvement in your depression.

Testing page PID=13

5. The side effects vary depending on the type of antidepressant you take.

Rating criterion #20:

"Exercise can be effective – alone or as an adjunct to other treatments."

Testing page PID=18

6. One such study showed that while antidepressants were fairly quick at improving symptoms of depression, after 16 weeks of treatment, exercise was equally effective as antidepressants in reducing depression in patients suffering major depressive disorder.

Figure 5-5 Examples rated as criterion #1, #6 and #20 successfully

The first example in Figure 5-5 was recognized as criterion #1 by both the rule-based approach and machine learning approach. This success demonstrates that the computer

programs were able to successfully map text expressions to semantic concepts, including "SSRI" – "antidepressant", "treating" – "treat", and "depression" – "major depressive disorder". Similarly, sentence examples from No.2 to 5 were rated as criterion #6 by both automated approaches. These examples demonstrate that many other text variations of "antidepressants" such as "SNRIs" and "MAOIs" were also successfully recognized. In addition, these examples include two different ways for expressing the meaning of criterion #6. One says directly that side effects "***vary***" depending on antidepressants; the second indicates variation by a discussion of "***fewer/more***" side effects between antidepressants. In both cases, the rule-based approach and machine learning approach successfully identified that the sentences are in concordance with the rating criterion #6.

In addition, the No.3 example in Figure 5-5 shows how hypernym of key concepts is processed in the rule-based approach. A general principle in rule-based approach is that if a semantic tag extracted from a sentence is a hypernym of a key concept referred to by a rating criterion, this hypernym alone is not to be considered as a valid substitute for this concept for the purposes of pattern matching in order to prevent false positives. For example, "side effects vary depending on medicines" is different from "side effects vary depending on antidepressants" because "medicines" in the former does not necessarily refer to antidepressant medication. Clearly, the determination of the objects referred by a hypernym (i.e. dereferencing a hypernym) is context dependent. To take context into consideration, the solution in this study is to apply a shifting window to scan the text immediately before the current sentence and check whether the hypernym term (e.g. medicine) is used to refer to a hyponym concept (e.g. antidepressant) in previous sentences. If so, the matching algorithm

considers the hypernym term in the current sentence as a valid match corresponding to the semantic unit specified in the criterion pattern. In this study, sentences after the current sentence were not included into the shifting window because review of the training cases indicated that this approach did not yield accurate results..

Rating criterion #6:

"The side effect profile varies for different antidepressants."

Testing page PID=2

Sentence in context: (The context in the shifting window is in italic font)

7. "*Taking an antidepressant for at least 6 months after you feel better can help keep you from getting depressed again. If this is not the first time you have been depressed, your doctor may want you to take the medicine even longer.*

   Side effects

   Side effects may vary depending on the **medicine** you take, but common ones include stomach upset, loss of appetite, diarrhea, feeling anxious or on edge, sleep problems, drowsiness, loss of sexual desire, and headaches."

Figure 5-6 Using context information to improve semantic processing

Figure 5-6 shows the example No.7, in which the sentence being processed is in regular font and the context in shifting window is in italic font. The shifting window comprises a number of sentences immediately prior to the current one. Through scanning the shifting window (see example in Figure 5-6), the term "medicine" in the current sentence is found to be linked, as a hypernym, to another concept "antidepressant" in previous sentences. Hence the rule-based system considers 'medicine" in the current sentence to be substitutable by "antidepressant" and consequently identifies the current sentence as a positive case for rating criterion #6. For

the final testing of 31 web pages, the shifting window included the three sentences immediately previous to the one under consideration. This size was identified as optimal through empirical observation and tuning during the training phase. From the computation efficiency perspective, it is problematic to apply a shifting window of an extremely large size. Moreover, an over-sized shifting window could increase errors in dereferencing hypernyms. In this study, window sizes from 1 to 5 were tested. Figure 5-7 shows the relationship between the criteria identification and window size for criterion #1 during the training phase. Before the application of shifting window scanning, there were 6 criterion sentences missed by rule-based quality rating, because hypernyms (e.g. "medicine", "drug", etc.) in the sentence were not accepted as valid substitutes for "antidepressant" despite other matching pattern features. By scanning text in the shifting window, some of these missing cases were identified because the rating programs were able to confirm that the hypernym actually referred to "antidepressant" in the context. The number of missed criteria decreased as the window size increased. At the same time, however, scanning the shifting window also introduced a few false positive cases (i.e. non-criterion sentences were mistakenly identified as a criterion). One false positive was introduced when the window size was one, and the number of false positive increased with the growth of window size. The accuracy of the criterion identification (i.e. the proportion of sentences that were correctly identified) was the highest when the window size was set to be 3. As shown in Figure 5-7, this shifting window (size = 3) results in the greatest improvement in criterion identification, while resulting in relatively few false positive identifications. Therefore, shifting window size was configured to be 3 in the learned system.

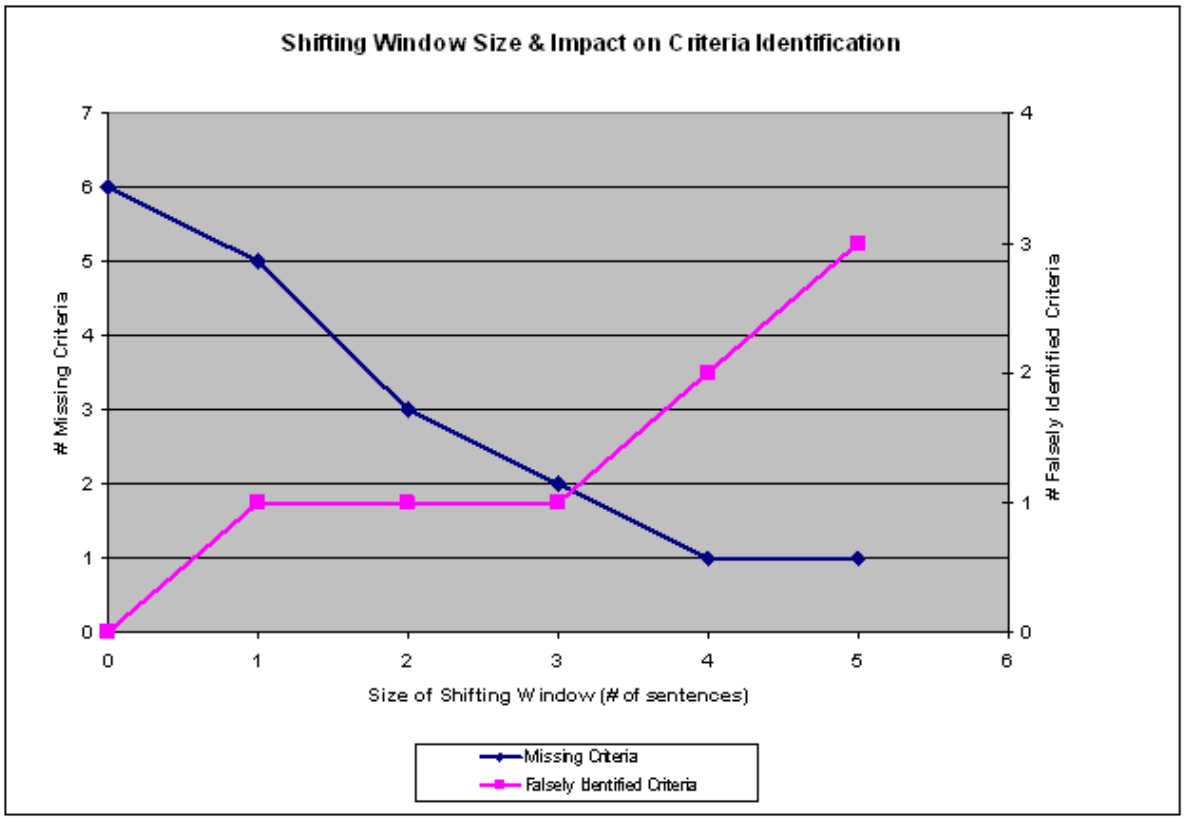

Figure 5-7 Size of shifting window

While most occurrences of the criteria were identified correctly by the automated approaches, there were also some difficult cases in which criterion-like sentences were not classified as a criterion (i.e. false negative) by automated approaches or a non-criterion sentence was mistakenly classified as an instance of a criterion (i.e. false positive).

5.3.2 False Negative Cases

Two types of false negative examples are listed in Figure 5-8. Example No.8 is for criterion #7. In this case, two consecutive sentences together convey the meaning of criterion #7, but neither of them alone covers the meaning of the criterion. Since the processing unit of the

current semantic processing algorithm is the sentence, expressions like this are still a challenge. The system needs to be enriched with more advanced processing capability to logically connect the meanings of multiple sentences. This study did not probe the analysis required to solve this type of issue, but efforts for making such improvement can be considered in future study.

In example No.9, the false negative decision on criterion #20 was related to a false identification of a sentence boundary. The text of sentence in question included a bullet list. The list was split into four independent sentences by the semantic tagging program, which is a common module for two automated rating approaches. Because none of these four "sentences" independently includes all the mandatory semantic meaning-bearing components for satisfying the classification patterns of criterion #20, the example No. 9 was not successfully identified. In spite of the above difficulties, the sentence classification performance of both the rule-based approach and machine learning approach is still good enough to identify most of the criteria recognized by human raters, as shown in section 5.2.

False Negative Cases

Rating Criterion #7

The choice of antidepressant should depend on individual patient factors (e.g. presence of co-morbid psychiatric or medical conditions, previous response to a particular drug, patient preference regarding the desirability of specific side-effects, concurrent drug therapy, suicidal risk)

Testing page PID=2

Sentences in context:

8. "Be sure your doctor knows about any other health conditions you have and any medicines you take regularly. This information can affect which antidepressant your doctor prescribes for you."

Rating Criterion #20

Exercise can be effective - alone or as an adjunct to other treatments.

Testing page PID=28

Plain text in paragraph:

9. "Regardless of whether you have mild or major depression, the following self-care steps can help:

* Get enough sleep.
* Follow a healthy, nutritious diet.
* Exercise regularly."

Figure 5-8 False negative examples rated by rule-based approach

5.3.3 False Positive Cases

Figure 5-9 contains two examples of false positive identifications made by either or both the rule-based approach and machine learning approach. Although it is identified as such by the machine learning approach (though not the rule-based approach), the human raters do not consider example No.10 to be a statement of criterion #1. This error appeared to arise because the sentence contains "your response to certain antidepressant", which causes the

sentence vector of the semantic representation to be very close to one or multiple positive training cases for criterion #1.

Another example (No.11) was incorrectly identified by the rule-based approach as an instance of criterion #6, while the machine learning approach correctly marked it as a negative case of this criterion. The semantic structure of this sentence looks almost the same as a frequently used expression - "*antidepressant*(A) has *more side effects* than antidepressant(B)", but with the replacement of "antidepressant(B)" by "natural alternatives". It caused false positive because the pattern definition was designed to increase recall by trading off on precision. In particular, although the semantic component corresponding to the antidepressant(B) was defined in the pattern for criterion #6, it was not mandatory in order to account for situations such as semantic ellipsis and resumptive pronouns. It has to be acknowledged that this simplified design is a compromise to accommodate the fact that the semantic analysis in this study did not deal directly with the challenge of "understanding" ellipsis and/or resumptive pronouns. Hence this simplification likely causes a false positive like example No.11. To obtain better performance in criterion identification, the semantic processing capability with respect to these features must be improved in future studies.

False Positive Cases:

Rating criterion #1:

"Antidepressant medication is an effective treatment for major depressive disorder."

Testing page PID=4 (rated as criterion #1 by machine learning)

10. The test, called the cytochrome P450, helps pinpoint genetic factors that influence your response to certain antidepressants (as well as some other medications).

Rating criterion #6:

"The side effect profile varies for different antidepressants."

Testing page PID=12 (rated as criterion #6 by rule-based classification)

11. It should be noted, though, that prescription medications carry with them far more side effects than their natural alternatives.

Figure 5-9 False positive examples rated by rule-based & machine learning approaches

5.4 Performance of Sentences Classification

This study takes an approach different from that used in previous research on automated quality rating. For example, Griffiths et al. (2005) rated the quality score of a website through a keyword analysis, and Wang and Liu (2007) developed an automatic indicator detection tool to collect indirect indicators for rating the web-based health care information quality. In the current study, the automated quality rating on depression treatment web pages is implemented based on a different foundation – i.e. sentence classification based on text semantics. For this reason, the performance evaluation in this study also includes an examination of the performance of sentence classification. Sentence-level evaluation provides a more detailed picture of the performance of the automated systems through a close look at the semantic classification results, rather than looking only at the assigned quality scores.

5.4.1 Performance of Rule-based Approach

The 31 testing web pages together contain 2677 sentences. When computer programs verify the sentences' concordance with a rating criterion, each sentence is classified relative to every criterion: that is, a decision is made whether or not the sentence represents an instance of each criterion. The values in the third and fourth columns in Table 5-4 list the number of sentences classified by rule-based approach as positive (Y) or negative (N) cases of each criterion. The first two rows in italic font under the table headers provide brief description of the contents of each cell. The last two rows for "Overall" are obtained by merging the results for all twenty-three criteria.

Table 5-4 Performance of sentence classification by rule-based approach

Rating Criteria	Human Classif-ication	Rule-based Classification (Y)	Rule-based Classification (N)	Recall	Precision	Accuracy
Criterion ID	*Y **	*True positives (TP)*	*False negatives (FN)*	*TP / (TP + FN)*	*TP / (TP + FP)*	*(TP + TN) / (TP + FN + FP + TN)*
	*N **	*False positives (FP)*	*True negatives (TN)*			
#1	**Y**	40	9	81.6%	85.1%	99.4%
	N	7	2621			
#2	**Y**	3	0	100.0%	42.9%	99.9%
	N	4	2670			
#3	**Y**	0	0	NA	NA	100.0%
	N	0	2677			
#4	**Y**	0	0	NA	NA	100.0%
	N	0	2677			
#5	**Y**	0	0	NA	NA	100.0%
	N	0	2677			
#6	**Y**	16	3	84.2%	84.2%	99.8%
	N	3	2655			
#7	**Y**	5	1	83.3%	55.6%	99.8%
	N	4	2667			

#8	**Y**	2	0	100.0%	100.0%	100.0%
	N	0	2675			
#9	**Y**	1	0	100.0%	50.0%	100.0%**
	N	1	2675			
#11	**Y**	6	2	75.0%	100.0%	99.9%
	N	0	2669			
#12-A	**Y**	9	2	81.8%	90.0%	99.9%
	N	1	2665			
#12-B	**Y**	10	3	76.9%	76.9%	99.8%
	N	3	2661			
#13-A	**Y**	4	0	100.0%	100.0%	100.0%
	N	0	2673			
#13-B	**Y**	2	0	100.0%	100.0%	100.0%
	N	0	2675			
#14	**Y**	14	4	77.8%	77.8%	99.7%
	N	4	2655			
#15	**Y**	2	0	100.0%	25.0%	99.8%
	N	6	2669			
#20	**Y**	9	3	75.0%	90.0%	99.9%
	N	1	2664			
Overall	**Y**	123	27	82.0%	78.3%	99.9%
	N	34	45325			

Note:
* The label 'Y' or 'N' in the second column stands for different human rating classification for each criterion (i.e. the class of sentences identified as criterion (Y) and those not identified as criterion (N)).

** The *accuracy* result for criterion #9 (i.e. 100%) is an approximation after being rounded up.

The last three columns in Table 5-4 are the measurements used in this study. *Precision* and *Recall* are typical performance measurements for evaluating the classification performance. The third measurement is *Accuracy*. Their definitions are given below.

- **Precision ***

  = the proportion of true positives (TP) over tested positives

  = TP / (TP + FP)

- **Recall ***

  = the proportion of true positives (TP) over actually positives

  = TP / (TP + FN)

- **Accuracy ***

  = the proportion of correctly identified sentences over all sentences

  = (FP + TN) / (TP + FN + FP + TN)

Precision and *recall* indicate the ability of the automated approaches to correctly identify positive instances of each criterion. The higher recall, the fewer actual criteria sentences go undetected (lower false negative rate). The higher precision, the fewer non-criterion cases are mistakenly identified as a criterion (lower false positive rate). For each web page in this study, the number of negative cases (i.e. FP + TN) for each criterion is far greater than the number of positive cases (i.e. TP + FN). The negative over positive ratios for all criteria is averagely 302:1. For single criteria, the negative over positive ratio ranges from the minimum of 54:1 (criterion #1) to the maximum of 2676:1 (criterion #9). Given the very low numbers of positive cases for each criterion, precision could be low because true positives (TP) can be easily overwhelmed by false positives (FP) even though only a very small portion of actually negative cases are mistakenly identified as positive. For this reason, the third performance indicator, i.e. *accuracy*, is used to take the skewed proportion of negative over positive cases into account and to indicate the combined capability of correctly identifying positive or negative of individual sentences.

Table 5-4 lists the values of performance indicators for the rule-based approach for each individual criterion and across criteria as well. This result was obtained based on the 31 testing web pages. Overall, the accuracy of classification results by rule-based approach is very high (> 99.4%). Recall ranges from 75% to 100%. The variation of recall across the criteria may be attributed to a variety of factors including the number of ways to paraphrase a specific criterion, the number of available positive training cases, and the coverage of different paraphrasing patterns in the training data. Average recall across all criteria is 82% and it was calculated based on the combination of every sentence case for all criteria, with each one equally weighted.

5.4.2 Performance of Machine Learning Approach

For the machine learning (i.e. Naïve Bayes) approach, the sentence classification performance is evaluated using the same measurements. The testing web pages are the same as those used in the testing of the rule-based approach, i.e. 31 web pages and in total 2677 sentences. As discussed in Chapter 4, due to the low number of positive instances of the criteria in the data corpus, the machine learning approach sentence classification performance was evaluated only for criteria #1, #6 and #12-B. Table 5-5 lists the performance of the machine learning approach for each individual criterion. For all three criteria, the recalls were above 84%. This shows that the machine learning approach effectively identified the sentences reflecting these three criteria, despite the natural language variations in criterion expression.

Table 5-5 Performance of sentence classification by machine learning approach

Rating Criteria	Human Classifi-cation	Machine Learning Classification (Y)	Machine Learning Classification (N)	Recall	Precision	Accuracy
#1	Y	42	7	85.7%	13.7%	89.9%
	N	263	2365			
#6	Y	16	3	84.2%	76.2%	99.7%
	N	5	2653			
#12-B	Y	11	2	84.6%	28.9%	98.9%
	N	27	2637			

Compared with the rule-based approach, the machine learning approach has slightly higher recall (85.7% vs. 81.6%, 84.2% vs. 84.2%, and 84.6% vs. 76.9%) on the same set of rating criteria (i.e. #1, #6 and #12-B). Precision, however, is much lower (13.7% vs. 85.1%, 76.2% vs. 84.2%, and 28.9% vs. 76.9%). Thus, the machine learning approach generates more false positives than the rule based approach. This may be due to the fact that the criterion matching patterns in the rule-based system not only take the semantic components into account, but also apply some constraints to screen out negative cases in order to increase the matching precision. For example, the constraint of proximity between a pair of semantic units (e.g. concepts, predications) in the rule-based approach enhances the relationship between semantic units in a single sentence, whereas the machine learning approach simply checks the co-occurrence of the semantic pairs. In addition, the rule-based approach also defines negation constraint in criterion matching patterns to decrease false positive rate by filtering out sentences which are anti-criterion cases. Therefore, the higher precision of the rule-based approach actually suggests that the machine learning approach still has room for improvement, since further studies can explore whether the machine learning classification model could include additional dimensions to represent the semantic and syntactic features

that assisted performance in the rule based approach. This future research could explore whether the inclusion of these additional features improves the performance of identifying rating criteria.

Overall, the testing in this thesis demonstrated that semantics-based quality rating (both rule-based and machine learning approaches) can produce quality score results comparable to human rating results. This is achieved by having computer programs to conduct shallow semantic analysis on each sentence in depression treatment web pages, and then use the semantic tag instance of training sentences to develop classifiers' capability to identify the sentences that are in concordance with the rating criteria. The identification of criterion-like sentences is treated as a binary classification of the semantic tag instance of sentences. The classification performance listed above attests to the efficacy of automatic quality score rating.

Chapter 6
Discussions, Conclusions, Contributions, and Future Work

This chapter briefly summarizes the methodology, discusses the quality rating results in depth, and presents directions for future research.

6.1 Summary of the Study

This study explores automated approaches for assessing the content quality of depression treatment web pages. The quality assessment is treated as a knowledge mining process, in which the target knowledge elements are the evidence-based depression treatment criteria which were developed from the best evidence in the depression treatment literature and published by the Centre for Evidence-based Mental Health at Oxford (CEBMH, 1998). The goal of the automated quality rating systems developed in this thesis is to identify sentences that convey these treatment criteria in depression treatment web pages. Two automated approaches are explored in this thesis: a rule-based approach and a machine learning (Naïve Bayes) approach. The rule-based quality rating system is applied to the whole rating criteria set, whereas the Naïve Bayes based quality rating system is tested on a subset of three criteria (i.e., #1, #6, and #12-B) as a proof of concept to demonstrate that semantics-based quality rating can be integrated with a fully automated algorithm. In each testing scenario, every web page is assigned a quality score equal to the number of unique treatment criteria identified by the respective rating method in each text. This process is automated using computer programs. The working procedures are: the computer programs read the text sentence by

sentence, and for each sentence determine whether it is in concordance with one of the rating criteria, incrementing the quality score by 1 for every unique criterion identified.

In this study, a semantics-based methodology is used to accomplish the above task. As discussed in Chapter 4, sentences in the raw text undergo a shallow semantic analysis in which expressions in English natural language are converted into semantic representations. Specifically, semantic tags for medical terms and noun phrases such as medical conditions, symptoms, medications or treatment names are represented using a controlled vocabulary (also called semantic concepts) in UMLS. Relationships between semantic concepts, which are often reflected in the use of verbs, adjectives and adverbs in the text, are processed by natural language processing modules (i.e. LVG) so that text variants such as inflectional variants of terms, possessives and synonyms are also transformed and normalized. Eventually, the semantic concepts extracted from each sentence, the semantic relationships between those concepts, and their position in the sentence are included in a semantic tag instance. These tags contain features for identifying whether a sentence is in concordance with a rating criterion.

After the features are extracted from the sentences and represented using semantic tags, the identification of the sentences that reflect a rating criterion is performed using two different methods: a rule-based classification approach and a Naive Bayes classification approach. The quality rating results generated by the two approaches are comparable to those of human raters: specifically, the quality scores assigned using the automated methods are strongly

correlated to human rating quality scores with statistical significance ($r=0.909$, $p<.001$ for the rule-based approach, and $r=0.841$, $p<.001$ or the Naïve Bayes approach).

Overall, the two research questions proposed in Chapter 2 have been answered by this study. First, a semantics-based approach was proposed for representing the sentence semantics at appropriate granularity, and the generated shallow semantic representation was demonstrated to be effective for supporting automatically rating information quality in depression treatment web pages. Although the UMLS based semantic parsing developed in this study did not try to extract the full semantics from the sentences, the approach was able to capture the semantic entities necessary for identifying sentences that reflect rating criteria. The success of this semantic parsing allowed the computer to conduct a "shallow" semantic analysis of sentences, and thus prepared a base for the automated classification of sentences based on semantics. Second, the testing and statistical analysis results show that semantics-based classification, i.e. rule-based and Naïve Bayes classifications were able to identify sentences in concordance with different rating criteria with reasonably satisfying accuracy. Due to the good performance of classification, both the rule-based rating system and the Naive Bayes based rating system generated quality scores that were strongly correlated to human rating quality scores. In summary, this study took a new approach to the assessment of the quality of health care information on the web, different from approaches used in previous studies that, for example, relied on indirect quality indicators such as web resource ownership. The quality ratings resulting from this new approach were demonstrated to be comparable to human ratings of information quality on depression treatment web pages.

6.2 Results and Discussion

As a result of the very low proportion of positive cases for a rating criterion (i.e., averagely 1 sentence that reflects a criterion for every 302 sentences that do not reflect this particular criterion) in the testing data, recall and accuracy better reflect classification performance than does precision. *Recall* is the proportion of actual criteria sentences identified as such and it indicates the likelihood of correctly identifying those sentences in the corpus that reflect the clinical guidelines; *Accuracy* represents the proportion of both criteria and non-criteria sentences that were correctly identified as each type. *Precision* is the percentage of true criteria sentences among all the sentences marked by the classification software as a criterion. Given that sentences that reflect a criterion make up only a very small proportion of the testing dataset, relatively low precision was anticipated. As discussed in Chapter 4, the quality rating results rely on the classification results of every sentence relative to each rating criterion, since the quality scores represent the number of unique criteria that are identified in the web page contents.

As shown in Table 5-4, for the rule-based approach across the 23 rating criteria, recall was 82.0%, and the overall accuracy was 99.9%. The rule-based rating of all rating criteria had overall precision around 78.3%. The quality score of a web page was assigned based on the classification result of all the sentences relative to every rating criterion. For this approach, the correlation between the rule-based quality scores (over all criteria) and human rating quality scores across the 31 testing pages was 0.909 ($p < .001$, $n = 31$).

The performance of the machine learning (Naïve Bayes) classification was evaluated using the same measurements. Unlike the rule-based classification, the testing for Naïve Bayes classifier did not cover all rating criteria. Instead, this testing focused on the three criteria that had a relatively large number of occurrences in order to effectively implement training and testing. In spite of this limited focus, the results still provide some indication of how well the shallow semantic analysis implemented in this study can be integrated with machine learning algorithm to accomplish automatic identification of sentences in concordance with rating criteria. Based on the results for the three selected rating criteria, the overall recall was 85.2%, accuracy was 96.2% and precision was 19%. Although the Naïve Bayes classification had low precision, there was still a strong Pearson correlation of 0.841 ($p < .001, n = 31$) between the automated quality scores obtained through the Naïve Bayes approach and the human rating quality scores (calculated for the three included items only). The fact that a high correlation was obtained while precision was low can be partially explained by the fact that the quality score reflects the number of unique criteria were identified in a web page. If the approach identifies 5 instances of criterion #1 on a page, the quality score is negatively influenced if and only if all instances are false positives. If even one of the instances is a true criterion sentence, the quality score will accurately reflect the presence of this rule, and will not be influenced by the additional false positives.

6.2.1 Analysis of Quality Score Rating Results

Although the rule-based system and the Naïve Bayes system use different sets of rating criteria for testing the quality assessment, in both cases the quality score produced by the automated rating was strongly correlated to the quality scores produced by human raters. In

all cases the quality scores for the depression treatment web pages were generated by verifying the concordance between the text content of the web pages and evidence-based rating criteria, rather than relying on any of the other types of quality indicators reviewed in Chapter 2. Specifically speaking, in the quality assessment process, each sentence in a depression treatment web page is automatically transformed into a shallow semantic representation, which is composed of semantic tags generated through semantic parsing and analysis. Using a rule-based or Naïve Bayes classification system, the semantic representation of every sentence is classified with respect to every rating criterion as either: a) an instance of that criterion or b) NOT an instance of that criterion. Thus, there are two classes for every criterion: Criterion sentences and NOT criterion sentences. The classification relative to a criterion is in fact a process of identifying those sentences that reflect the specific rating criterion. The results of this study suggest that this semantics-based approach could be a new promising way to rate health care information quality.

Obviously, a successful quality rating method must be able to address all relevant rating criteria for a health care information topic such as depression treatment. In this study, the shallow semantic representation of sentences acts as a basis for comparing the web page text content with depression treatment rating criteria. The method for generating the shallow semantic representations worked well to support content-based quality rating. The testing of the rule-based system was conducted using the whole set of rating criteria. The automatically generated quality scores were strongly correlated ($r=0.909$, $p<0.01$, $n=31$) with human rating results. The satisfying results suggested the effectiveness of semantics-based approach in two aspects. First, the shallow semantic representation generated for a sentence was generically

effective for classification tasks relative to all rating criteria, instead of being tailored for each classification task respectively. The development of the semantic tagging process in this study was independent from the rating criteria in that no processing was customized to deal with any specific rating criterion and its unique content and concepts. Second, the semantic representation of sentences was at a level of granularity appropriate to meet the purpose of quality rating. Of course, the generalizability of the method still needs to be proved via quality rating test on other health conditions.

The Naïve Bayes based quality rating was implemented in this study on a subset of the rating criteria primarily as a proof of concept. It should be noted that the design of Naïve Bayes based rating in this study is generically applicable to all depression treatment rating criteria. Due, however, to restrictions inherent in the data (low incidence of many criteria on a random web page providing few positive cases for machine-based learning) and the time and resource limit on organizing human rating over larger size of data, the testing of Naïve Bayes based rating was conducted on #1, #6, and #12-B only. Nevertheless, the significance of the Naïve Bayes testing is that it demonstrated that the employed semantic analysis techniques and the semantic processing programs developed in this study can be successfully integrated with automatic algorithms (specifically in this case Naïve Bayes classification). Measured by the same performance indicators used in the evaluation of rule-based rating system, the Naïve Bayes approach demonstrated good performance in creating quality ratings based on criteria #1, #6 and #12-B. To conclude, the testing results indicated that the Naïve Bayes quality scores can be valid indicators of quality as reflected in the scores assigned by human raters for the criteria #1, #6 and #12-B. In a comparison between the rule-based rating and Naïve

Bayes based rating on the same criteria set, the results of the rule-based rating were slightly superior to the results from the Naïve Bayes approach ($r = .841$, $p < .001$ for the Naïve Bayes approach compared to $r = .897$, $p < .001$ for the same three criteria under the rule-based approach). The performance advantage of the rule-based rating system may be attributed to the knowledge engineering involved in that approach, which potentially contributed more features such as proximity constraint between co-occurring semantic units to the classification model. This may also suggest that adding more syntactic and semantic features to the Naïve Bayes classification model might improve performance. This could be explored and tested in a future study of semantics-based quality rating.

6.2.2 Applicability of Semantics-based Quality Rating to Other Health Conditions

The subject scope selected in this study was the depression treatment knowledge domain. But what will be required if the semantics-based rating approach is applied to the treatment of a different medical condition or other non-treatment health subjects? As has been explained in Chapter 4, the transformation from text to shallow semantic representation provided the foundation for implementing semantics-based quality rating. In order to guarantee the generalizability of the semantic representation of the health care web documents, this study selected the UMLS tool set to implement semantic tagging. The UMLS is designed and maintained by the U.S. National Library of Medicine. It consists of knowledge sources (e.g. Metathesaurus) to support the mapping and translating of biomedical concepts and a set of software tools (e.g. LVG, MMTx) to support the parsing of natural language text and return of semantic concepts (National Library of Medicine, 2009). Since the UMLS knowledge sources contain more than 60 families of biomedical controlled vocabularies including MeSH

and SNOMED CT (National Library of Medicine, 2009), it is expected that most if not all health conditions and subjects could be effectively processed with the methods used in this study to tag depression treatment web documents.

Regarding semantic classification, the rule-based classification implemented in this study employed human knowledge engineering to generate the classification rules. As introduced in Section 4.3.1, in order to establish classification rules with semantic concepts and their relations that are necessary for identifying sentences reflecting a rating criterion, the knowledge engineer studied the expressions of positive sentences identified by human raters in the training samples. Therefore, one limit of the rule-based classification approach proposed in this study is that it will require manual knowledge engineering efforts every time a classification system needs to be built for a new set of rating criteria. But it should also be clarified that the methodology for knowledge engineering in this study is generic since the constraints (e.g. co-occurrence, position relations) that were employed for specifying classification rules apply across health subjects.

In contrast, the Naïve Bayes classification proposed in this study has some advantages in saving human effort. From training to testing, every process in Figure 4-7 (2a) is handled by computer programs. Changes in the health subject or medical condition are not expected to limit the employment of the Naïve Bayes classification method, because in this approach semantic tags are converted into dimensions in the vector space model and therefore the medical concepts are transparent to the learning of the classifier.

6.2.3 Comparison of Quality Assessment Results

In contrast to the semantic processing and analysis used in the current research, previous research in the automatic assessment of health care information quality (Griffiths et al., 2005; Hawking et al. 2007; Tang et al. 2009) used a keyword analysis approach to rate the quality score of 30 depression treatment websites.

In Griffiths et al. (2005) study, the Pearson correlation between the quality scores resulting from the keyword approach and the human rating results was also high ($r=0.850, p < .001, n = 30$). Thus, the keyword based approach seems to be effective for automatically rating information quality. Nonetheless, there are some advantages to the approaches used in the current study over the keyword approach to automatic rating.

First, in the keyword-based approach, Griffiths et al. (2005) used the text documents from the training websites to train a learned relevance query and a learned quality query. The learned queries are composed of terms for which the term frequency distribution discriminates positive training documents from negative training documents, weighted using Robertson-Sparck Jones formula (Robertson & Jones, 1976). The resulting queries are used to derive quality scores for the testing websites based on the similarity between website documents and the learned queries. Thus, the site score generated by the keyword approach is a scaled value based on similarity ratio in the range from 0 to 1, and the score does not represent the number of rating criteria that are endorsed by a website. In contrast, the scores generated by the semantics-based approaches in the current research are more meaningful. Each score indicates the number of the rating criteria that are identified in the text being assessed.

Second, due to the scoring nature of the keyword-based approach, it is difficult to explain the reason for a particular quality score. Provided with two quality scores (say 0.8 versus 0.81), users may likely feel unconfident with understanding the true difference between two scores and selecting one source over the other. In contrast, using the semantics-based approach proposed in this study, the integer-valued score credited to any given web page can be justified by listing the criteria that the web page text contains and where they are located. The transparency of the scoring process in the semantics-based approach could increase user confidence in the quality assessment results.

Third, with assistance from the semantics-based rating approach, health care information users and stakeholders can have greater insight into the content strengths and weakness of web pages. Although two web pages can receive the same quality rating, the information published on these two pages may cover different criteria, and the semantics-based rating approach described in this thesis can provide this detail. This information is critical to a better understanding of the strengths and weaknesses of each resource. As shown in the case examples in Chapter 5, the semantics-based approach is able to provide detailed rating reports that display the information coverage of web pages. Inclusion of such a knowledge coverage profile as a part of the quality assessment results could provide assistance to information users over and above the information provided by the quality score.

6.2.4 Comparison of Automated Quality Rating Approaches

The semantics-based quality assessment approach in this study is different from previous keyword-based approach and approaches using non-content quality indicators (e.g. Smith,

2002; Frické et al., 2005; etc.) not only in quality rating results, but also in other aspects, including the indicators utilized, the features being analyzed, and the web sources being rated. Table 6-1 provides a summary of the contrasts.

Table 6-1 Comparison of automated assessment rating approaches

Compared Items		**Semantics-based Approach**	**Keyword frequency based Approach**	**Approaches using indirect quality indicators**
Quality Standard/Indicators		Evidence-based treatment criteria (provided by CEBMH)	Evidence-based treatment criteria (provided by CEBMH)	website accountability standards (e.g. disclosure of authority, sponsorship, interest conflict, etc.); webmetric indicators (e.g. hyperlinks, traffic); others
Rating object (Granularity)		Web pages	Websites	Websites
Features being processed		semantic concepts, semantic relationships, combinations patterns	key terms and frequency	non-content metadata
Vulnerability to term spamming		Lower	Higher	N/A
Rating Results	**Meaning of Quality Score**	Indicates the number of qualified content indicators	Similarity to learned queries	Indicates the number of qualified non-content indicators
	Provision of explanation on quality score	Yes (listing out treatment criteria covered by the text)	No	Yes(non-content metadata of the website)
	Ability of indicating the content strength	Yes (assist users to know the aspect of content has strength)	No	N/A
Capability of dealing with dynamic content update on assessed objects		Yes (re-run content matching)	Yes	No or not timely enough
Knowledge Subject		Depression Treatment	Depression Treatment	Different health care domains
Extendibility to different subject domain		Supposed to be	Supposed to be	Yes

First, in terms of quality indicators, rating approaches using indirect indicators do not assess content. Instead, they rate quality based on an examination of the disclosure of relevant meta-information on the website, such as authorship and sponsorship. The meta-information adopted for quality rating can usually be derived from some practical standards, for example HONcode (2012), which defines a set of standards to which health care website are expected to comply. Eysenbach et al. (2002) identified the 25 most frequently used indirect indicators based on a systematic review of 79 distinct studies of Internet health care information quality assessment. In a more recent study, Griffiths et al. (2005) tested using a type of webmetric metadata, i.e. Google PageRank of a web site as indicators of information quality. The PageRank is developed by Google founders Brin and Page (1998) to evaluate the reputation of a web page based on computation of web citations linking to the page. As reviewed in Chapter 2, although a correlation between some web site meta-information and website content quality was identified in some previous studies (e.g. Chen et al., 2000; Fallis & Fricke, 2002; Griffiths & Christensen, 2005; etc.), other studies (e.g. Martin-Facklam et al. 2003; Griffiths et al., 2005; Khazaal et al., 2012; etc.) found that the correlations were not statistically significant. Such differences seem unavoidable due to the fact that the indirect quality indicators say nothing directly about web site content.

In contrast, content-based rating approaches use content indicators to evaluate information quality. Particularly, the semantics-based approach and the keyword-based approach rely on well-established evidence-based clinical guidelines as content-based quality indicators. Such evidence-based clinical guidelines are widely available not only for depression treatment, but also other medical conditions. The “evidence-based medicine” concept (EBMWG, 1992) was

originated by a group of McMaster scholars in 1992. Since that time, evidence-based clinical guidelines have been published and advocated by authoritative medical groups and institutes in different countries. Some examples of such resources include the *Agency for Healthcare Research and Quality* website (AHRQ, 2012) created by U.S. Department of Health & Human Services, and the *National Institute for Health and Clinical Excellence* (NICE, 2012) in the United Kingdom. The authoritative guidelines provide a sufficient source for rating criteria to implement the semantics-based quality assessment. The current study was conducted within a specific knowledge domain (i.e. depression treatment); future studies can explore different knowledge domains including treatment for other medical conditions and/or different types of health care information such as symptom diagnosis in order to prove the generalizability of the methodology and the robustness of the system developed in this study.

Second, the rating approaches in Table 6-1 are used to assess the quality of information at different granularity levels in terms of rating objects. Websites are the object of quality rating in most approaches using indirect indicators simply because most indicators are meta-information at the site level. Griffiths' keyword-based approach (2005) was implemented and tested at the website level as well. In contrast, the semantics-based approach proposed in this study works at the web page level. As online information seekers most often retrieve pages from web search engines, the provision of quality score at page level granularity has advantages in providing assistance to search engine users in their decisions regarding web page retrieval. However, the semantics-based approach can also be applied to rating the quality of a health care website, by aggregating the quality scores of all contained pages. The aggregation can be performed by counting the number of unique criteria identified across the

website. Of course, other ways of aggregating scores, e.g. normalization by the total count of web pages, can be examined in future studies.

Third, although the keyword approach by Griffiths et al. (2005) and the semantics-based approaches in the current study both use the text content for automated assessment, they are different in terms of the features being processed. Griffiths' approach focuses on keywords and their frequency distribution, while the approaches discussed in this study involve intensive semantic processing of the text, examining not only semantic concepts (comparable to keywords), but also semantic relationships between concepts and the patterns that are used to put these semantic units together to form a sentence. Griffiths et al. (2005) acknowledge that the keyword approach could be compromised if publishers use spamming methods for optimizing their automatic quality scores. Theoretically, the semantics-based approach is less vulnerable to a spamming attack. A 'pile of keywords' spam would not easily cause the semantics-based rating system to identify criteria by mistake, because a criterion-like sentence must not only contain the key semantic concepts but also organize the semantic units in specific patterns in order to express meaning effectively.

Last but not least, the reality of dynamic web page content is a challenge for the quality assessment of health care information on the web. Typical non-content indicator approaches rely on relatively stable characteristics such as authorship and sponsorship that do not always change at the same pace as the content. Hence, it is likely that the quality scores resulting from these approaches will remain the same even after content updates such as the correction of mis-information or addition of new information. Undoubtedly, content-based quality

rating approaches have the advantage in this respect. Every time the semantics-based quality rating programs conduct a quality assessment, the quality score reflects the current content.

Based on the above discussion and comparison, it can be concluded that the semantic quality assessment approach in this study is able to automate the quality rating process, and the rating performance is strongly correlated with human ratings. In addition, the semantics-based approach has certain advantages over previously proposed solutions as listed in Table 6-1. We expect that the semantics based automatic quality rating approach can contribute to the improvement of web-based health care information service and consumption. For example, one bright prospect for its application is to integrate the automated quality rating results into the web search engine results. Quality rating results could be displayed in a side bar for each retrieved web page returned by search engines, allowing online search users to rely not only on content relevance but also the information quality to decide which page they want to navigate. The semantics-based quality rating could be provided either as an independent third-party service, or as a plug-in function under the search engine umbrella.

6.3 Limitations and Future Work

There are some limitations in this study. First, automated quality rating relies on pre-existing evidence-based quality criteria, and the approach cannot be applied when such criteria are not available. As the prevalence of evidence-based medical practice increases, evidence-based clinical guidelines are becoming widespread and are being developed for a wider range of medical conditions. These evidence-based clinical guidelines are candidate sources for generating quality criteria. Second, as introduced in Chapter 2, the content quality has four

important properties: correctness, comprehensiveness, bias-free and currency. In this study, the quality score generated by the semantics-based rating approach directly reflects the first two properties since 1) the presence of an identified rating criterion (i.e. evidence-based health practice guideline) is an indicator of information correctness of web documents; 2) the quality score is determined by how many different rating criteria were covered by the web page content, i.e. the comprehensiveness of the text content. However, it needs to be acknowledged that the current design of the semantics-based quality rating approach has not yet dealt with offsetting quality score with information which goes directly against the rating criteria. Thus, this type of anti-criteria sentence, if contained in a web page, does not contribute negatively to web page quality score. In addition, the semantics-based approach proposed in this study does not directly examine content currency and content bias. These two criteria could be partially addressed by the underlying clinical guidelines, in that these evidence-based clinical guidelines are likely to be bias-free and up to date. If a rating criterion has been derived from the latest evidence-based practice, the identification of that criterion on a web page attests to the currency of that information on the web page. At the same time, the information on the web page could be mixed with other health care information that is outdated, and this outdated information will not be recognized by the system as such due to the lack of relevant criteria, and thus will have no effect on the quality score. Similarly, this limitation also exists for the examination of content bias. It should be noted though that such issues could also influence human rating if that rating is performed using the same rating criteria. Third, it has been acknowledged in Chapter 3 that the approach in this study is not designed to rate web pages containing complex non-text formats. For example, the proposed approach cannot assess the quality of information contained in

images, and if a web page contains tables, the content inside the table may not be properly processed. In addition, multimedia web sources, such as audios and videos, cannot be evaluated by this approach unless the information content is converted into text format by speech recognition or other technologies. Fourth, the semantics-based approach proposed in this study is focused on the "factual" aspect of content analysis, while the analysis of attitude and affect in the text is not covered. However, in recent years an increasing amount of web content, particularly in the social media zone such as Facebook and Twitter, is rich in subjective opinions (e.g. Kelly, 2009), rather than facts. For example, a patient may tweet his/her personal feeling about a health condition experience, a health care professional may share stories about a treatment trial and his/her subjective comments through micro-blogging. Quality assessment based on such "opinion-based" information will have to rely more on an analysis of sentiment in the content, including attitude expressions, writer's certainty, and writing stylistic features such as forms of reference, tenses, types of evidential language, etc. In the last few years, investigation on such topics has gradually emerged as a new research subject, i.e., sentiment analysis, and has achieved some success (e.g. Shanahan et al., 2006; Rubin & Vashchilko, 2012). Exploration of using state-of-the-art sentiment analysis technology for quality assessment of web health care information will be a necessary complement to the semantics-based approach proposed in this study.

This study is to our knowledge the first attempt to use a semantics-based approach (i.e. through the comparison of text content with rating criteria) to automatically rate health care information quality on the web. In the current research on depression treatment web pages, the rule-based rating system and the Naïve Bayes based rating system each produced quality

scores strongly correlated to human rating results. Theoretically speaking, the techniques and tools employed in this study, including transformation from text to semantic tags, classification methods for identifying sentences in concordance with rating criteria, and the UMLS resource, can be applied to process text in other biomedical sub-domains, as there is no domain-specific design in this approach. This study was conducted based on a solid foundation of previous research (CEMBH, 1998; Griffiths & Christensen, 2002), in which previous researchers have made advances in efforts to summarize and refine the evidence-based depression treatment guidelines into a set of one-sentence statements so that each guideline statement could be easily converted into rating criteria in the current study. In addition, it has been demonstrated previously (Griffiths & Christensen 2002; Griffiths & Christensen, 2005; Griffiths et al., 2005) that the human rating quality scores generated based on these specific evidence guidelines are highly correlated with subjective rating performed by health care professionals; moreover, the generation of quality scores based on these guidelines has been demonstrated to have reasonably high inter-rater reliability. These factors were important to the successful quality rating achieved in this study on depression treatment web pages. Future studies can be conducted to verify whether this approach can succeed in rating information quality of other health conditions.

6.4 Overall Conclusion

This study proposed a semantics-based quality rating approach for automatically assessing the evidence-based content quality of health care information on the web. This approach is demonstrated to be successful in rating the content quality of depression treatment web pages. The rule-based rating system, which adopts manually extracted patterns, produced

quality scores with strong correlation to human rating quality scores. The quality scores automatically rated by the Naïve Bayes rating system were also strongly correlated to human rating quality scores in the rating test on a smaller set of rating criteria. The strong correlations show that the automatically generated quality scores can be valid indicators of the quality of depression treatment web pages. In comparison to previous research, the rating result produced by the semantics-based approach has the advantage of providing more detailed insights regarding the quality of the web source content, and thus has potential to offer health care information consumers more support in information search and navigation. If the results of this thesis are replicable and generalizable to other health conditions, this semantics based approach could add significant value to the quality assessment practice of health care information on the web.

References

Ademiluyi, G., Rees, C.E. & Sheard, C.E. (2003). Evaluating the reliability and validity of three tools to assess the quality of health care information on the Internet. *Patient Education and Counseling*, *50(2)*, 151-155.

Aronson, A. (1994). *Comparison of LVG and MetaMap functionality*. Retrieved from http://skr.nlm.nih.gov/papers/references/LVG-MetaMap.comparison.pdf

About.com (2009). *Medical search engines*. Retrieved from http://websearch.about.com/od/enginesanddirectories/tp/medical.htm

Anderson, K. A., Nikzad-Terhune, K. A., & Gaugler, J. E. (2009). A systematic evaluation of online resources for dementia caregivers. *Journal of Consumer Health on the Internet, 13(1)*, 1-13. doi:10.1080/15398280802674560

AHCPR Depression Guideline Panel (1993). *Depression in Primary Care. Volume 2. Detection and Diagnosis. Clinical Practice Guideline, No 5.* Rockville, MD. U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research. AHCPR Publication No. 93-0551. April 1993.

AHRQ (2011). Clinical practice guidelines archive. Retrieved from http://www.ahrq.gov/clinic/cpgarchv.htm

Baker, C. F. & Hiroaki S. (2003). The FrameNet data and software. *The companion volume to the Proceedings of 41st annual meeting of the association for computational linguistics. Ed. Yuji Matsumoto. 2003*. (pp. 161-164).

Barnes, C., Harvey, R., Wilde, A., et al. (2009). Review of the quality of information on bipolar disorder on the Internet. *Australian and New Zealand Journal of Psychiatry, 43*, 934-945.

Bath, P. A. & Bouchier, H. (2003). Development and application of a tool designed to evaluate web sites providing information on Alzheimer's disease. *Journal of Information Science, 29(4)*, 279–297.

Baujard, V., Boyer, C., & Geissbhler, A. (2011). Evolution of health web certification through the HONcode experience. *Studies in Health Technology and Informatics, 169*, 53-57

Berland, G. K., Elliot, M. N., Morales, L. S., Algazy, J. I., Kravitz, R. L., Broder, M. S., … McGlynn, E. (2001). Health care information on the Internet accessibility, quality, and readability in English and Spanish. *The Journal of the American Medical Association, 285(20)*, 2612-2621.

Bouchier, H. & Bath, P. A. (2003). Evaluation of web sites that provide informationon Alzheimer's disease. *Health Informatics Journal, 9(1)*, 17–31.

Billsus, D. & Pazzani, M. (1999). A hybrid user model for news story classification. *Proceeding UM '99 Proceedings of the seventh international conference on User modeling. Banff, Canada. June, 1999*. (pp. 99-108). Springer-Verlag New York.

Boyer, C. & Geissbuhler, A. (2005). A decade devoted to improving online health care information quality. *Studies in Health Technology and Informatics,116,* 891-896

Bopp, R.C. & Smith, L.E. (2000). *Reference and information services: An introduction.* Libraries Unlimited. 3Rev Ed edition

Brin, S. & Page, L. (1998). Anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*, 107-117

Burkell, J. (2004). Health care information seals of approval: what do they signify? *Information, Communication & Society*, *7(4)*, 491-509.

CEBMH, (1998). *CEBMH depression guideline treatment*. Retrieved from http://web.archive.org/web/20040426143952/http://cebmh.warne.ox.ac.uk/cebmh/guidelines/depression/treatment.html

Contact a Family & Information Society Research and Consultancy Group. (2007). Judge: web sites for health. Retrieved from http://www.judgehealth.org.uk/consumer_guidelines.htm

Charnock, D., Shepperd, S., Needham, G. & Gann, R. (1999). DISCERN: an instrument for judging the quality of written consumer health care information on treatment choices. *Journal of Epidemiology Community Health*, *53*, 105-111.

Chen, L.E., Minkes, R.K., Langer, J.C. (2000) Pediatric surgery on the Internet. *Journal of Pediatric Surgery*, *35*, 1179-1182.

Chen, J., Huang, H., Tian, S. & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications. 36(3),* 5432-5435

Chen P., Barrera, A., Rhodes, C. (2010). Semantic analysis of free text and its application on automatically assigning ICD-9-CM Codes to patient records. *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference 7-9 July 2010* (pp. 68-74).

comScore. (2011). *comScore releases June 2011 U.S. search engine rankings*. Retrieved from http://www.comscore.com/Press_Events/Press_Releases/2011/7/comScore_Releases_June_2011_U.S._Search_Engine_Rankings

Crespo, J. (2004). Training the health care information seeker: quality issues in health care information Web sites. *Library Trends*, *53(22)*, 360-374

Crocco, A.G., Villasis-Keever, M. & Jadad, A.R. (2002). Analysis of cases of harm associated with use of health care information on the internet. *Journal of the American Medical Association*, *287(21)*, 2869-2871.

Deleger L, Grouin C, Zweigenbaum P. (2010). Extracting medical information from narrative patient records: the case of medication-related information. *Journal of American Medical Informatics Association*, *17*, 555–558.

Dunlop, M. (1997). The effect of accessing nonmatching documents on relevance feedback. *ACM Transaction on Information Systems, 15(2)*, 137 – 153.

Eysenbach, G. & Diepgen, T.L. (1998). Towards quality management of medical information on the Internet: evaluation, labeling, and filtering of information. *British Medical Journal*, *317(1)*, 496-502.

Eysenbach G, Köhler C, Yihune G, Lampe K, Cross P, Brickley D. (2001). A framework for improving the quality of health care information on the world-wide-web and bettering public (e-)health: the MedCERTAIN approach. *Medical Information. 10(Pt 2)*:1450-1454

Eysenbach, G. Powell, J., Kuss, O. & Sa, E. (2002). Empirical studies assessing the quality of health care information for consumers on the world wide web. *Journal of the American Medical Assosication*, *287(20)*: 2691-2700.

Eysenbach, G. & Köhler, C. (2004). Health-related searches on the Internet. *Journal of the American Medical Assosication*, *291(24)*, 2946.

Evidence-Based Medicine Working Group. (1992). Evidence-based medicine: a new approach to teaching the practice of medicine. *The Journal of the American Medical Association. 268(17)*, 2420-2425.

eWebMarketing, (2009). *Search Engine Optimisation & Marketing*. Retrieved from http://www.ewebmarketing.com.au/

Fallis, D & Frické M. (2002). Indicators of accuracy of consumer health care information on the Internet: a study of indicators relating to information for managing fever in children in the home. *Journal of the American Medical Informatics Association*, *9(1),* 73-79.

Frické, M. & Fallis, D. (2002). Verifiable health care information on the Internet. *Journal of Education for Library and Information Science*, *43(4)*, 246-253.

Frické, M. & Fallis, D. (2004). Indicators of accuracy for answers to ready reference questions on the Internet. *Journal of the American Society for Information Science and Technology*, *55(3)*, 238-245.

Frické, M., Fallis, D., Jones, M. & Luszko, G. M. (2005). Consumer health care information on the Internet about carpal tunnel syndrome: Indicators of accuracy. *The American Journal of Medicine. 118*, 168-174.

Graham, L., Tse, T. & Keselman, A. (2006). Exploring user navigation during online health care information seeking. *AMIA Annu Symp Proc. 2006*. 299-303.

Griffiths, K.M. & Christensen, H. 2000. Quality of web based information on treatment of depression: cross sectional survey. *British Medical Journal, 321*, 1511-1515.

Griffiths, K.M. & Christensen, H. (2002). The quality and accessibility of Australian depression sites on the World Wide Web. *The Medical Journal of Australia*, *160*, 97-104.

Griffiths, K.M. & Christensen, H. (2005) Website quality indicators for consumers. *Journal of Medical Internet Research, 7(5)*:e55. Retrieved from http://www.jmir.org/2005/5/e55/

Griffiths, K.M., Tang, T.T., Hawking, D. and Christensen, H. 2005. Automated assessment of the quality of depression websites. *Journal of Medical Internet Research, 7(5)*:e59. Retrieved from http://www.jmir.org/2005/5/e59/

Grishman, R. & Sundheim, B. (1996). Message understanding conference- 6: A brief history**.** In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, *Kopenhagen, 1996*, (pp. 466–471).

Guan, H., Zhou, J. & Guo, M. (2009). A class-feature-centroid classifier for text categorization. *Proceedings of the 18th international conference on World Wide Web. Madrid, Spain, April 20-24, 2009* (pp. 201-210). ACM. NY.

Han, E. H. & Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results. *Principles of data mining and knowledge discovery, 4th European conference, PKDD 2000, Lyon, France, September 13-16, 2000* (pp. 424-431).

Hawking, D., Tang, T., Sankaranaravana, R., Griffiths, K. Craswell, N. & Bailey, P. (2007). Towards higher quality health search results: Automated quality rating of depression websites. *In Proceedings of Medinfo 2007 Workshop on* "*Models of trust for health websites*". *August, 2007*.

HealthInsite. (2011). *The HealthInsite concept*. Retrieved from http://www.healthinsite.gov.au/content/internal/page.cfm?ObjID=0001AC1D-0806-1D2D-81CF83032BFA006D

HealthLinkBC. (2011). *About HealthLink BC*. Retrieved from http://www.healthlinkbc.ca/aboutprogram.stm/

Hesse, B.W., Nelson, D.E., Kreps, G.L., Croyle, R.T., Arora, N.K. & Rimer, B.K. (2005). The impact of the Internet and its implications for health care providers: findings from the first health care information national trends survey. *Archive Internal Medicine*, *165*: 2618-2624.

HONcode (2012). *HONcode: Principles – Quality and trustworthy health care information*. Retrieved from http://www.hon.ch/HONcode/Conduct.html

Humphreys, B., Lindberg, D., Schoolman, H., and Barnett, G. (1998). The Unified Medical Language System: An informatics research collaboration. *Journal of the American Medical Informatics Association, 5*, 1-11

Internet News (2003). *September 2003 Internet Usage Stats*. Retrieved from http://www.internetnews.com/stats/article.php/3096031/September+2003+Internet+Usage+Stats.htm

Internet World Stats. (2012) *North America Internet usage statistics, population and telecommunications reports*. Retrieved from http://www.internetworldstats.com/stats14.htm

Kelly, R. (Ed.) (2009). *Twitter Study – August 2009*. Retrieved from http://www.pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf

Khazaal, Y., Chatton, A., Zullino, D. & Khan, R. (2012). HON label and DISCERN as content quality indicators of health-related websites. *The Psychiatric quarterly, 83(1)*, 15-27.

Khazaal, Y., Chatton, A,, Cochand, S., Coquard, O., Fernandez, S., Khan, R., … Zullino, D. Brief DISCERN, six questions for the evaluation of evidence-based content of health-related websites. *Patient education and counseling, 77(1)*, 33-7.

Kiley, R. (2002). Does the Internet harm health? Some evidence does exist that the Internet harms health. *British Medical Journal*, *323(7331)*, 328-329.

Kim, H. Howland, P. & Park H. (2005). Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research, 6(1)*. p37-53

Koenemann, J. (1996). Supporting interactive information retrieval through relevance feedback. *Proceedings of the CHI'96 Conference*, 49-50.

Kunst, H. Groot, D., Latthe, P. Latthe, M. & Khan, K.S. (2002). Accuracy of information on apparently credible websites: Survey of five common health topics. *British Medical Journal*, *321(7337)*: 581-582.

Leman, H. (2008). *The Top 10 Health Search Engines of 2008*. Retrieved from http://www.altsearchengines.com/2008/12/29/the-top-10-health-search-engines-of-2008/

Liu, T., Chen Z., Zhang, B. Ma, W., Wu, G. (2004). Improving text classification using local latent semantic indexing. *ICDM'04, Fourth IEEE International Conference on Data Mining*, 2004. 162-169.

Martin-Facklam, M., Kostrzewa, M., Martin, P. & Haefili, W.E. (2003). Quality of drug information on the World Wide Web and strategies to improve pages with poor information quality: An intervention study on pages about sildenafil. *British Journal of Clinial Pharmacology*, *57(1)*, 80-85.

McCallum, A. & Nigam, K. (1998). A comparison of Event Models for Naïve Bayes Text Classification. *IN AAAI-98 Workshop on Learning for Text Categorization*. (pp.41-48). AAAI Press.

McCray, A.T., Srinivasan, S. & Browne, A.C. (1994). Lexical methods for managing variation in biomedical terminologies. *the Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, *1994*, (pp. 235-239).

Medline Plus. (2011). *Medline plus guide to healthy web surfing*. Retrieved from http://www.nlm.nih.gov/medlineplus/healthywebsurfing.html

Mednar, (2009). *Mednar – deep web medical search*. Retrieved from http://mednar.com/mednar/about.html

Nadeau, D. & Sekine, S. (2007) A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1):3–26.

National Library of Medicine (2009). *UMLS reference manual*. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK9676/

Nielsen, (2010). *Top U.S. search stes for May 2010*. Retrieved from http://blog.nielsen.com/nielsenwire/online_mobile/top-u-s-search-sites-for-may-2010/

NHS. (2011). *NHS choices about us – Editorial policy*. Retrieved from http://www.nhs.uk/aboutNHSChoices/aboutnhschoices/Aboutus/Pages/Editorialpolicy.aspx

NICE (2012). *About NICE guidance*. Retrieved from http://www.nice.org. uk/guidance /index.jsp

Peterson, G., Aslani, P. & Williams, K.A. (2003). How do consumers search for and appraise information on medicines on the Internet? A qualitative study using focus groups. *Journal of Medical Internet Research*, *5(4)*, e33. Retrieved June 25, 2007, from http://www.jmir.org/2003/4/e33/index.htm

Pew Internet and American Life Project. (2003). *Internet health resources*. Retrieved from http://www.pewinternet.org/~/media//Files/Reports/2003/PIP_Health_ Report_July_2003.pdf

Pew Internet and American Life Project. (2006). *Online health search 2006*. Retrieved from http://www.pewinternet.org/~/media/Files/Reports/2006/PIP_Online_ Health_2006.pdf

Pew Internet and American Life Project. (2011). *The social life of health care information, 2011*. Retrieved from http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx

Podichetty, V.K., Booher, J., Whitfield, M. & Biscup, R.S. (2006). Assessment of internet use and effects among healthcare professionals: a cross sectional survey. *Postgraduate Medical Journal, 82*:274-279

Portney, L.G. & Watkins, M.P. (2000) *Foundations of Clinical Research: Applications to Practice. 2nd ed*. Upper Saddle River, NJ: Prentice Hall Health.

Reeve, L. H. (2007). *Semantic annotation and summarization of biomedical text*. Philadelphia, PA, Drexel University. ISBN: 978-0-549-13037-6

Rindflesch, T. & Aronson, A. (2002). *Semantic processing for enhanced access to bomedical knowledge*. Retrieved from http://skr.nlm.nih.gov /papers/references/semwebapp.5a.pdf

Robertson, S. E. & Jones, K.S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 129–146.

Rubin, V.L. & Vashchilko, T. (2012) Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. *In Proceedings of the 13th Conference of the European Chapter for the Association for Computational Linguistics: Computational Approached to Deception Detection Workshop, Avignon, France, April 23, 2012* (pp. 97-106).

Sahami, M., Dumais, S., Heckerman, D., Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. *AAAI'98 Workshop on Learning for Text Categorization*.

Salton, G. & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*. *41(4)*, 288-297.

Schapire, R. E. & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning, 39(2/3)*,135-168.

Schneider, K. M. (2003). A comparison of event models for Naïve Bayes anti-spam email filtering. *Proceeding EACL '03 Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, Budapest, Hungray April 12-17, 2003* (pp. 307-314). Association for Computational Linguistics, Stroudsburg, PA, USA

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*, 1-47.

Shanahan, J. G., Qu, Y. & Wiebe, J., (Eds.). (2006). *Computing attitude and affect in text: Theory and applications (The information retrieval series)*. New York, NY: Springer.

Silberg, W. M., Lundberg, G. D. & Musaccio, R. A. (1997) Assessing, controlling, and assuring the quality of medical information on the Internet: caveat lector et viewor--let the reader and viewer beware. *Journal of the American Medical Association*, *277*, 1244–1245.

Smith, A. (2002). *Evaluation of information sources*. *The World-Wide Web Virtual Library*. Retrieved from http://www.vuw.ac.nz/staff/alastair_smith/avaln/evaln.htm

Tang, T., Hawking, D., Sankaranarayana, R., Griffiths, K., Craswell, N. (Eds.) (2009, April). Quality-oriented search for depression portals. In *ECIR '09 Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Toulouse, France, April 6-9, 200,*. (pp.637-644). Berlin, Heidelberg. Springer.

UMLS. (2009). *UMLS – Metathesaurus release statistics*. Retrieved from http://web.archive.org/web/20090925122534/http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

URAC. (2007). *URAC promoting quality health care*. Retrieved from http://www.urac.org/

Vaughan, L. & Thelwall, M. (2005). A modeling approach to uncover hyperlink patterns: the case of Canadian universities. *Information Processing and Management*, *41*, 347-359

Wang, M. & Liu, C. (2009). Class selection based iterative supervised latent semantic indexing for text categorization. In *ICIECS 2009. International Conference on Information Engineering and Computer Science,* 2009. p1-4.

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

Wang, M., Nie, J. Zeng, X. (2005). A latent semantic classification model. *CIKM'05 Proceedings of the 14th ACM international conference on Information and knowledge management*. 261-262.

Wiener, E., Pederson, J. O., & Wiegend, A. S. (1995). A neural network approach to topic spotting. In *Symposium on Document Analysis and Information Retrieval,* Las Vegas, NV (pp. 317–332).

WEKA. (2011). *Weka 3: Data Mining Software in Java*. Retrieved from http://weka.wikispaces.com/Primer

Witten, I. H., Frank, E. & Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques. (Third Edition)*. Morgan Kaufmann.

Appendices

Appendix A: DISCERN

An instrument for judging the quality of written consumer health care information on treatment choices Funded by the British Library

For further information please contact: Sasha Shepperd University of Oxford Division of Public Health and Primary Health Care Institute of Health Sciences Old Road Headington Oxford OX3 7LF

Section 1

IS THE PUBLICATION RELIABLE?

1. Are the aims clear?

No		Partially		Yes
1	2	3	4	5

*Hint: Look for a clear indication at the beginning of the publication of * what it is about * what it is meant to cover (and what topics are meant to be excluded) * who might find it useful If the answer to Question 1 is 'No', go directly to Question 3*

2. Does it achieve its aims?

No		Partially		Yes
1	2	3	4	5

Hint: Consider whether the publication provides the information it aimed to as outlined in Questin 1

3. Is it relevant?

No		Partially		Yes
1	2	3	4	5

*Hint: Consider whether * the publication addresses the questions that readers might ask * recommendations and suggestions concerning treatment choices are realistic or appropriate*

4. Is it clear what sources of information were used to compile the publication (other than the author or producer)?

No		Partially		Yes
1	2	3	4	5

*Hint: * Check whether the main claims or statements made about treatment choices are accompanied by a reference to the sources used as evidence (e.g. a research study or expert opinion) * Look for a means of checking the sources used such as a bibliography reference list or the addresses of the experts or organizations quoted*

Rating note: In order to score a full '5' the publication should fulfill both hints. Lists of additional sources of support and information (Q.7) are not necessarily sources of evidence for the current publication

5. Is it clear when the information used or reported in the publication was produced?

No		Partially		Yes
1	2	3	4	5

*Hint: Look for * dates of the main sources of information used to compile the publication * date of any revisions of the publication (but not dates of reprinting) * date of publication (copyright date)*

Rating note: The hints are placed in order of importance - in order to score a full '5' the dates relating to the first hint should be found

6. Is it balanced and unbiased?

No		Partially		Yes
1	2	3	4	5

*Hint: Look for * a clear indication of whether the publication is written from a personal or objective point of view * evidence that a range of sources of information was used to compile the publication (e.g. more than one research study or expert) * evidence of an external assessment of the publication Be wary if * the publication focuses on the advantages or disadvantages of one particular treatment choice without reference to other possible choices * the publication relies pi— manly on evidence from single cases (which may not be typical of people with this condition or of responses to a particular treatment) * the information is presented in a sensational, emotive or alarmist way*

7. Does it provide details of additional sources of support and information?

No		Partially		Yes
1	2	3	4	5

Hint: Look for suggestions for further reading or for details of other organisations providing advice and information about the condition and treatment choices

8. Does it refer to areas of uncertainty?

No		Partially		Yes
1	2	3	4	5

*Hint: * Look for discussion of the gaps in knowledge or differences in expert opinion concerning treatment choices * Be wary if the publication implies that a treatment choice affects everyone in the same way (e.g. 10000 success rate with a particular treatment)*

Section 2

HOW GOOD IS THE QUALITY OF INFORMATION ON TREATMENT CHOICES?

N.B. The questions apply to the treatment (or treatments) described **in the publication**. Self-care is considered a form of treatment throughout this section.

9. Does it describe how each treatment works?

No		Partially		Yes
1	2	3	4	5

Hint: Look for a description of how a treatment acts on the body to achieve its effect

10. Does it describe the benefits of each treatment?

No		Partially		Yes
1	2	3	4	5

Hint: Benefits can include controlling or getting rid of symptoms, preventing recurrence of the condition and eliminating the condition - both short-term and long-term

11. Does it describe the risks of each treatment?

No		Partially		Yes
1	2	3	4	5

Hint: Risks can include side effects, complications and adverse reactions to treatment - both short-term and long-term

12. Does it describe what would happen if no treatment is used?

No		Partially		Yes
1	2	3	4	5

Hint: Look for a description of the risks and benefits of postponing treatment, of watchful waiting (i.e. monitoring how the condition progresses without treatment) or of permanently forgoing treatment

13. Does it describe how the treatment choices affect overall quality of life?

No		Partially		Yes
1	2	3	4	5

*Hint: Look for * description of the effects of the treatment choices on day-to-day activity* description of the effects of the treatment choices on relationships with family, friends and carers*

14. Is it clear that there may be more than one possible treatment choice?

No		Partially		Yes
1	2	3	4	5

*Hint: Look for * a description of who is most likely to benefit from each treatment choice mentioned, and under what circumstances * suggestions of alternatives to consider or investigate further (including choices not fully described in the publication) before deciding whether to selector reject a particular treatment choice*

15. Does it provide support for shared decision-making?				
No		Partially		Yes
1	2	3	4	5

Hint: Look for suggestions of things to discuss with family, friends, doctors or other health professionals concerning treatment choices

Section 3

OVERALL RATING OF THE PUBLICATION

16. Based on the answers to all of the above questions, rate the overall quality of the publication as a source of information about treatment choices:				
Low *Serious or extensive shortcomings*		Moderate *Potentially important but not serious shortcomings*		High *Minimal shortcomings*
1	2	3	4	5

Appendix B: Data Sampling

Table B-1 Depression treatment web page samples (Whole Set)

ID	URL
1	http://www.mayoclinic.com/health/depression/DS00175/DSECTION=treatments-and-drugs
2	http://www.emedicinehealth.com/depression/page5_em.htm
3	http://www.emedicinehealth.com/depression/page6_em.htm
4	http://www.emedicinehealth.com/depression/page7_em.htm
5	http://www.finddepressiontreatment.com/
6	http://health.usnews.com/articles/health/healthday/2009/05/27/stigma-keeps-teens-from-depression-treatment.html
7	http://www.mentalhealth.com/rx/p23-md01.html
8	http://www.effexorxr.com/depression-anxiety-treatment.aspx
9	http://www.waldenbehavioralcare.com/depression_treatment.asp
10	http://www.healthlinkbc.ca/kbase/topic/special/hw30709/sec1.htm
11	http://www.depressioncenter.org/treatments/cbt.asp
12	http://www.depressioncenter.org/treatments/meds.asp
13	http://www.depressioncenter.org/treatments/default.asp
14	http://www.hypnosisdownloads.com/cat/depression-treatment.html
15	http://www.medpagetoday.com/Psychiatry/Depression/14476
16	http://www.nlm.nih.gov/medlineplus/ency/article/003213.htm
17	http://www.healthlinkbc.ca/kbase/topic/special/hw30709/sec9.htm
18	http://www.healthlinkbc.ca/kbase/topic/special/hw30709/sec10.htm
19	http://www.webmd.com/anxiety-panic/features/alternative-depression-treatment-risks
20	http://en.wikipedia.org/wiki/Clinical_depression
21	http://www.mayoclinic.org/depression/treatment.html
22	http://www.healthlinkbc.ca/kbase/topic/special/hw30709/sec11.htm
23	http://www.sciencedaily.com/releases/2008/11/081130201928.htm
24	http://online.wsj.com/article/SB119128055574245655.html?mod=health_home_stories
25	http://www.medicalnewstoday.com/articles/81578.php
26	http://www.healthlinkbc.ca/kbase/topic/special/hw30709/sec12.htm
27	http://www.depression-guide.com/
28	http://psychologyinfo.com/depression/treatment.htm
29	http://www.iampanicked.com/anxiety-articles/depression-treatment-methods.htm
30	http://www.nimh.nih.gov/health/publications/depression/complete-index.shtml
31	http://depressionandanxietyhelp.com/depression-treatment.html
32	http://www.helpguide.org/mental/medications_depression.htm
33	http://helpguide.org/mental/treatment_strategies_depression.htm
34	http://safedepressiontreatment.com/

35	http://depressiontreatmentworks.org/
36	http://thedepressiontreatment.com/antidepressants/index.htm
37	http://www.depression.com/treatment_tips.html
38	http://depression-assistance.com/2006/07/30/depression-treatment/
39	http://depressiontreatment.net/
40	http://www.depression-guide.com/treatment-of-depression.htm
41	http://www.depression-help-treatment.com/depression-medication.html
42	http://www.depression-treatment-help.com/depression-treatment/depression-treatment.htm
43	http://www.depression-treatment-help.com/
44	http://health.yahoo.com/depression-treatment/depression-treatment-overview/healthwise--aa25747.html
45	http://www.depressiontreatmenthelp.org/depression_treatment.php
46	http://www.psychologyinfo.com/depression
47	www.psychologyinfo.com/depression/treatment.htm
48	http://mayoclinic.com/health/depression/DS00175/DSECTION=treatments-and-drugs
49	http://psychcentral.com/lib/2006/depression-treatment/all/1/
50	http://www.drweil.com/drw/u/ART00696/depression-treatment
51	http://www.mayoclinic.com/health/treatment-resistant-depression/DN00016
52	http://www.healthlinkbc.ca/kbase/dp/topic/ty6745/dp.htm
53	http://en.wikipedia.org/wiki/Depression_and_natural_therapies
54	http://au.reachout.com/find/articles/depression-management-and-treatment-options
55	http://www.aboutourkids.org/families/disorders_treatments/az_disorder_guide/depression/treatment
56	http://www.cancer.gov/cancertopics/pdq/supportivecare/depression/Patient/page4
57	http://www.cancer.gov/cancertopics/pdq/supportivecare/depression/Patient/91.cdr
58	http://www.healthlinkbc.ca/kbase/topic/detail/drug/hw29716/detail.htm
59	http://www.nimh.nih.gov/health/topics/child-and-adolescent-mental-health/antidepressant-medications-for-children-and-adolescents-information-for-parents-and-caregivers.shtml
60	http://www.aboutourkids.org/families/disorders_treatments/az_disorder_guide/depression/questions_answers
61	http://www.healthline.com/adamcontent/adolescent-depression
62	http://www.healthlinkbc.ca/kbase/topic/detail/drug/hw29398/detail.htm
63	http://www.healthline.com/adamcontent/depression-elderly
64	http://www.healthline.com/adamcontent/major-depression-with-psychotic-features
65	http://www.mayoclinic.com/health/depression-treatment/AN00685
66	http://www.omhrc.gov/templates/news.aspx?ID=627661
67	http://www.aboutourkids.org/families/disorders_treatments/az_disorder_guide/depression/treatment
68	http://www.nlm.nih.gov/medlineplus/news/fullstory_82699.html
69	http://www.ahrq.gov/clinic/epcsums/deprsumm.htm
70	http://www.oas.samhsa.gov/2k8/depression/depressionTX.cfm

71	http://mednar.com/mednar//mednar/link.html?collectionCode=HEL-IMPRO&searchId=fdf05ca7-40a4-4e18-af8e-3c5fe82088ce&type=RESULT_EMAIL&redirectUrl=https%3A%2F%2Fwww.acponline.org%2Fatpro%2Ftimssnet%2Fimages%2Fbooks%2Fsample%2520chapters%2FPsychCh05.pdf
72	http://www.nlm.nih.gov/medlineplus/news/fullstory_82699.html
73	http://www.healthlinkbc.ca/kbase/topic/detail/drug/hw29806/detail.htm
74	http://www.mayoclinic.com/health/alternative-medicine-side-effects/MY00682
75	http://www.healthlinkbc.ca/kbase/topic/detail/drug/hw29535/detail.htm
76	http://www.healthlinkbc.ca/kbase/dp/topic/zx3018/dp.htm
77	http://www.mayoclinic.com/health/depression-and-aging/MY00259
78	http://www.healthline.com/adamcontent/major-depression
79	http://www.healthline.com/galecontent/depression-1
80	http://www.medicinenet.com/script/main/art.asp?articlekey=52498
81	http://www.personalmd.com/news/a1996080501.shtml
82	http://familydoctor.org/online/famdocen/home/common/mentalhealth/treatment/012.html
83	http://www.nami.org/Template.cfm?Section=About_Treatments_and_Supports&template=/ContentManagement/ContentDisplay.cfm&ContentID=7952
84	http://netwellness.org/healthtopics/depression/depressiontreatment.cfm
85	http://www.healthycookingrecipes.com/articles-submit/david-mcevoy/depression-treatment.html
86	http://www.med.umich.edu/depression/treatment.htm
87	http://yourtotalhealth.ivillage.com/depression-treatment-hasnt-worked.html
88	http://www.realmentalhealth.com/alternatives/pleasant_activities.asp
89	http://www.dukenews.duke.edu/2000/02/mm_depressiontreatment.html
90	http://www.namiscc.org/Recovery/2002/NonDrugDepressionTreatments.htm
91	http://www.holisticonline.com/Remedies/Depression/dep_treatment_behavioral.htm
92	http://www.wdxcyber.com/psychotherapy.html
93	http://pibhs.uams.edu/Depression/Depression_treatment.asp
94	http://www.suicideandmentalhealthassociationinternational.org/depressiontreat.html
95	http://endoflifecare.tripod.com/huntdiseasefaqs/id49.html
96	http://allcare.net/s2/anxiety.php
97	http://www.healthlinkbc.ca/kbase/as/tb1954/why.htm
98	http://www.healthlinkbc.ca/kbase/topic/major/ty4640/descrip.htm
99	http://www.healthlinkbc.ca/kbase/topic/major/ty4640/trtover.htm
100	http://www.webmd.com/depression/understanding-depression-treatment
101	http://www.webmd.com/depression/guide/depression-treatment-options
102	http://www.webmd.com/depression/postpartum-depression/understanding-postpartum-depression-treatment
103	http://www.webmd.com/depression/psychotherapy-treat-depression
104	http://www.webmd.com/depression/treating-depression-medication
105	http://www.webmd.com/depression/pediatric-prozac
106	http://www.webmd.com/depression/news/20080221/hope-may-take-time-after-depression

107	http://www.webmd.com/depression/continuum-care-treatment-resistant-depression
108	http://www.webmd.com/depression/guide/treatment-resistant-depression-psychotherapy
109	http://www.webmd.com/depression/experimental-treatments-depression
110	http://www.healthlinkbc.ca/kbase/dp/topic/ty6886/dp.htm
111	http://sh-print.healthwise.net/moh/print/PrintTableOfContents.aspx?token=moh&localization=en-ca&version=Q3_09&docId=tb1939
112	http://sh-print.healthwise.net/moh/print/PrintTableOfContents.aspx?token=moh&localization=en-ca&version=Q3_09&docId=tb1954
113	http://www.healthlinkbc.ca/kbase/topic/detail/drug/zp2718/detail.htm
114	http://www.healthlinkbc.ca/kbase/topic/major/tn9653/descrip.htm
115	http://www.webmd.com/depression/medication-options
116	http://www.webmd.com/depression/news/20081201/which-kids-need-antidepressants
117	http://www.webmd.com/depression/ssris-myths-and-facts-about-antidepressants
118	http://www.healthlinkbc.ca/kbase/topic/major/tn9653/trtover.htm
119	http://www.webmd.com/depression/news/20080303/fda-oks-new-antidepressant-pristiq
120	http://www.healthlinkbc.ca/kbase/topic/major/tn9653/drugtrt.htm
121	http://www.webmd.com/depression/guide/optimizing-depression-medicines
122	http://www.webmd.com/depression/guide/chronic-illnesses-depression
123	http://www.healthlinkbc.ca/kbase/topic/major/tn9653/othertrt.htm
124	http://www.healthlinkbc.ca/kbase/topic/major/tn9653/hometrt.htm
125	http://www.webmd.com/depression/guide/depresssion-support
126	http://www.webmd.com/depression/guide/depression-chronic-pain
127	http://www.webmd.com/depression/news/20090602/coping-skills-may-reduce-teen-depression
128	http://www.healthlinkbc.ca/kbase/topic/detail/drug/tn9670/detail.htm
129	http://www.healthlinkbc.ca/kbase/topic/detail/drug/tn9677/detail.htm
130	http://www.webmd.com/depression/adjusting-life-recovery
131	http://www.webmd.com/depression/guide/st-johns-wort
132	http://www.webmd.com/depression/guide/alternative-therapies-depression
133	http://www.webmd.com/depression/news/20080226/therapy-medication-switch-for-teen-depression
134	http://www.webmd.com/depression/guide/sexual-problems-and-depression
135	http://www.anxiety-and-depression-solutions.com/wellness_concerns/depression/depression_treatment.php
136	http://www.zoloft.com/depr_treatment.aspx
137	http://www.wdxcyber.com/psychotherapy.html
138	http://depression.emedtv.com/depression/depression-treatment.html
139	http://www.antidepressantsfacts.com/1995-12-Antonuccio-therapy-vs-med.htm
140	http://www.genf20.com/depression-treatment.html
141	http://depressiontreatment.net.au/
142	http://www.bayridgetreatmentcenter.com/depression.html

143	http://www.bodyhealthsoul.com/depression.htm
144	http://www.ayushveda.com/health/depression.htm
145	http://health.yahoo.com/depression-treatment/should-i-take-medications-to-treat-depression/healthwise--ty6745.html;_ylt=AkghCk5Z4QCEPGEl1CGHvw_EtcUF
146	www.familydoctor.org/handouts/587.html
147	http://www.nlm.nih.gov/medlineplus/antidepressants.html
148	http://effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=sg&DocID=10&ProcessID=7
149	http://familydoctor.org/online/famdocen/home/common/mentalhealth/treatment/012.printerview.html
150	http://www.nimh.nih.gov/health/publications/mental-health-medications/complete-index.shtml
151	http://www.consumerreports.org/health/best-buy-drugs/antidepressants.htm
152	http://www.mayoclinic.com/print/antidepressants/HQ01069/METHOD=print
153	http://familydoctor.org/online/famdocen/home/common/mentalhealth/treatment/045.printerview.html
154	http://www.fda.gov/Drugs/DrugSafety/InformationbyDrugClass/ucm096305.htm
155	http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm095980.htm
156	http://familydoctor.org/online/famdocen/home/common/mentalhealth/treatment/904.printerview.html
157	http://www.nimh.nih.gov/health/topics/child-and-adolescent-mental-health/antidepressant-medications-for-children-and-adolescents-information-for-parents-and-caregivers.shtml
158	http://www.mayoclinic.com/print/antidepressants/MH00059/METHOD=print
159	http://www.nimh.nih.gov/health/publications/depression-easy-to-read/index.shtml
160	http://www.nlm.nih.gov/medlineplus/tutorials/depression/mh019103.pdf
161	http://womenshealth.gov/faq/depression.cfm
162	http://www.fda.gov/ForConsumers/ByAudience/ForWomen/ucm118515.htm
163	http://www.nimh.nih.gov/health/publications/depression/how-is-depression-detected-and-treated.shtml
164	http://www.mayoclinic.com/print/depression/DS00175/DSECTION=all&METHOD=print
165	http://jama.ama-assn.org/cgi/reprint/300/18/2202.pdf
166	http://familydoctor.org/online/famdocen/home/common/mentalhealth/treatment/882.printerview.html
167	http://www.nami.org/Template.cfm?Section=About_Treatments_and_Supports&template=/ContentManagement/ContentDisplay.cfm&ContentID=7952
168	https://healthyontario.com/ConditionDetails.aspx?disease_id=43
169	http://www.healthlinkbc.ca/kbase/as/tb1939/what.htm
170	http://apahelpcenter.org/articles/article.php?id=52
171	http://www.mayoclinic.com/print/psychotherapy/MY00186/METHOD=print&DSECTION=all
172	http://www.healthlinkbc.ca/kbase/nci/ncicdr0000062806.htm
173	http://nccam.nih.gov/health/stjohnswort/sjw-and-depression.htm
174	http://www.mayoclinic.com/print/depression-and-exercise/MH00043/METHOD=print
175	http://www.intelihealth.com/IH/ihtIH/WSIHW000/8596/35226/363129.html?d=dmtContent
176	http://www.mayoclinic.com/print/clinical-depression/AN01057/METHOD=print

177	http://mentalhealth.samhsa.gov/publications/allpubs/ken98-0049/default.asp
178	http://www.lupus.org/webmodules/webarticlesnet/templates/new_learnliving.aspx?articleid=2256&zoneid=527
179	http://www.annals.org/cgi/reprint/149/10/I-56.pdf
180	http://www.nia.nih.gov/HealthInformation/Publications/depression.htm
181	http://www.nimh.nih.gov/health/publications/older-adults-depression-and-suicide-facts-fact-sheet/index.shtml
182	http://www.nimh.nih.gov/health/publications/women-and-depression-discovering-hope/how-is-depression-diagnosed-and-treated.shtml
183	http://www.healthyminds.org/Document-Library/Brochure-Library/Lets-Talk-Facts-Depression.aspx
184	http://www.mentalhealthamerica.net/go/information/get-info/depression/depression-in-teens
185	http://www.mayoclinic.com/print/depression-treatment/AN00685/METHOD=print
186	http://www.nimh.nih.gov/health/publications/treatment-of-children-with-mental-disorders/index.shtml
187	http://kidshealth.org/parent/emotions/feelings/understanding_depression.html
188	http://www.nlm.nih.gov/medlineplus/ency/article/000945.htm
189	http://www.blackdoginstitute.org.au/public/depression/treatments/psychological.cfm
190	http://www.blackdoginstitute.org.au/public/depression/treatments/index.cfm
191	http://www.healthlinkbc.ca/kbase/as/ug4845/actionset.htm
192	http://www.betterhealth.vic.gov.au/bhcv2/bhcarticles.nsf/pages/Depression_and_exercise
193	http://www.betterhealth.vic.gov.au/bhcv2/bhcarticles.nsf/pages/Depression_coping_and_recovering
194	http://www.healthlinkbc.ca/kbase/as/ug4814/actionset.htm
195	http://www.healthlinkbc.ca/kbase/as/tn9165/actionset.htm
196	http://www.healthlinkbc.ca/kbase/nci/ncicdr0000062739.htm
197	http://www.depressionservices.org.au/treatments/exercise-2.html
198	http://www.healthlinkbc.ca/kbase/as/uf9919/actionset.htm
199	http://www.nhs.uk/pathways/depression/pages/treatment.aspx
200	http://www.nhs.uk/conditions/postnataldepression/pages/treatment.aspx
201	http://www.nhs.uk/conditions/depression/pages/treatment.aspx

Frequency Distribution of the Quality of Web Page Samples

The quality score of the web page samples were divided into five bands and the frequency distribution is shown in Figure B-1.

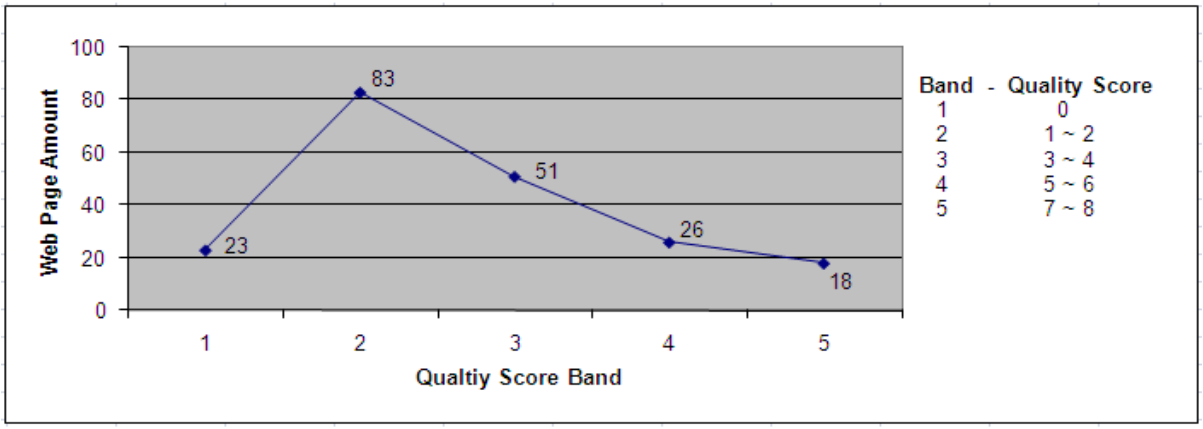

Figure B-1 Depression treatment web page samples (whole set)

URLs of Testing Web Pages

A stratified random sampling was conducted to select 31 web pages from the whole data set as testing pages. The URLs are listed in Table B-2.

Table B-2 The URL of the testing web page samples

TestID	URL
1	http://www.nimh.nih.gov/health/publications/depression/how-is-depression-detected-and-treated.shtml
2	http://sh-print.healthwise.net/moh/print/PrintTableOfContents.aspx?token=moh&localization=en-ca&version=Q3_09&docId=tb1939
3	http://www.mayoclinic.com/print/depression/DS00175/DSECTION=all&METHOD=print
4	http://www.mayoclinic.com/print/antidepressants/HQ01069/METHOD=print
5	http://www.webmd.com/depression/understanding-depression-treatment
6	http://www.helpguide.org/mental/medications_depression.htm
7	http://www.healthlinkbc.ca/kbase/topic/detail/drug/hw29398/detail.htm
8	http://www.emedicinehealth.com/depression/page7_em.htm
9	http://www.nia.nih.gov/HealthInformation/Publications/depression.htm
10	http://www.healthlinkbc.ca/kbase/topic/detail/drug/tn9677/detail.htm
11	http://www.depression-treatment-help.com/depression-treatment/depression-treatment.htm
12	http://www.depressiontreatmenthelp.org/depression_treatment.php

13	http://www.webmd.com/depression/treating-depression-medication
14	https://healthyontario.com/ConditionDetails.aspx?disease_id=43
15	http://psychcentral.com/lib/2006/depression-treatment/all/1/
16	http://www.healthlinkbc.ca/kbase/nci/ncicdr0000062806.htm
17	http://health.yahoo.com/depression-treatment/depression-treatment-overview/healthwise--aa25747.html
18	http://www.depressionservices.org.au/treatments/exercise-2.html
19	http://jama.ama-assn.org/cgi/reprint/300/18/2202.pdf
20	http://www.nlm.nih.gov/medlineplus/ency/article/000945.htm
21	http://familydoctor.org/online/famdocen/home/common/mentalhealth/treatment/012.html
22	http://www.healthlinkbc.ca/kbase/topic/major/ty4640/trtover.htm
23	http://www.mentalhealthamerica.net/go/information/get-info/depression/depression-in-teens
24	http://mednar.com/mednar//mednar/link.html?collectionCode=HEL-IMPRO&searchId=fdf05ca7-40a4-4e18-af8e-3c5fe82088ce&type=RESULT_EMAIL&redirectUrl=https%3A%2F%2Fwww.acponline.org%2Fatpro%2Ftimssnet%2Fimages%2Fbooks%2Fsample%2520chapters%2FPsychCh05.pdf
25	http://www.nimh.nih.gov/health/publications/treatment-of-children-with-mental-disorders/index.shtml
26	http://www.webmd.com/depression/news/20080221/hope-may-take-time-after-depression
27	http://www.webmd.com/depression/guide/sexual-problems-and-depression
28	http://www.nlm.nih.gov/medlineplus/ency/article/003213.htm
29	http://www.medpagetoday.com/Psychiatry/Depression/14476
30	http://depressionandanxietyhelp.com/depression-treatment.html
31	http://www.dukenews.duke.edu/2000/02/mm_depressiontreatment.html

Table B-3 URLs of web resources for collecting data samples

Web Source Name	Portal URL	Web Page Collection Time
Generic Search Engine		
Google	http://www.google.com/	May, 2009
Yahoo! Search	http://ca.search.yahoo.com/	May, 2009
Microsoft Bing Search	http://www.bing.com/	May, 2009
Ask.com	http://www.ask.com/	May, 2009
AOL	http://search.aol.com/aol/webhome	May, 2009
Medical Search Engine		
OmniMedicalSearch	http://www.omnimedicalsearch.com/	May, 2009
HealthFinder	http://www.healthfinder.gov/	May, 2009
HealthLine	http://www.healthline.com/	May, 2009
MedNar	http://mednar.com/mednar/	May, 2009
WebMD	http://www.webmd.com/search/	May, 2009
Health Care Web Portals		
Medline Plus	http://www.nlm.nih.gov/medlineplus/depression.html	May, 2009
HealthlinkBC	http://www.healthlinkbc.ca	May, 2009
HealthInsite	http://healthinsite.gov.au	May, 2009
National Health Service	http://www.nhs.uk/Pages/HomePage.aspx	May, 2009

Appendix C: Evidence-based Depression Treatment Guideline & Rating Criteria

1. Rating Criteria Used in (Griffiths & Christensen, 2005)

Evidence-Based Rating Scale for Human Raters (Copied from (Griffith & Christensen, 2005))

The evidence-based rating scale was developed from statements in the treatment section of *A systematic guide for the management of depression in primary care* published by the Centre for Evidence-based mental health, Oxford.

1. Antidepressant medication is an effective treatment for major depressive disorder.
2. Antidepressants are all equally effective.
3. The effectiveness of antidepressants is around 50 to 60%.
4. Full psychosocial recovery can take several months.
5. Drop out rate is same for different antidepressants.
6. The side effect profile varies for different antidepressants.
7. The choice of antidepressant should depend on individual patient factors (e.g. presence of co-morbid psychiatric or medical conditions, previous response to a particular drug, patient preference regarding the desirability of specific side-effects, concurrent drug therapy, suicidal risk)
8. Antidepressants are not addictive.
9. A trial of 6 weeks at full dose is needed before a drug can be considered to have failed and another tried.
10. A second-line drug should probably be from a different class of antidepressant.
11. Once improved continuation treatment at the same dose for at least 4-6 months should be considered.
12. Discontinuation syndrome may occur with abrupt cessation of any antidepressant so antidepressants should not be stopped suddenly. Where possible antidepressants should be withdrawn over a 4 week period, unless there are urgent medical reasons to stop the drug more rapidly. [To score 1, need to make general points that abrupt cessation can cause discontinuation syndrome and that antidepressants should not be stopped suddenly]
13. St John's Wort appears to be as effective as tricyclic antidepressants and causes fewer side effects, but little is known about any long term adverse effects.
14. Cognitive therapy can be an effective treatment for depression.
15. Cognitive behaviour therapy is at least as effective as drug treatment in mild-to-moderate depression.
16. Cognitive behaviour therapy may be valuable for people who respond to the concept of Cognitive behaviour therapy, prefer psychological to antidepressant treatment, have not responded to antidepressant therapy. [Score 1 if mention at least one of these]
17. Problem-solving may be effective for depression.
18. [Generic] counselling is probably no more effective than treatment as usual from the GP for depression.
19. Written information (usually based on a cognitive model of depression) can improve mild-to-moderate depression. [Score 1 if cognitive model]
20. Exercise can be effective - alone or as an adjunct to other treatments.

For each item, score 1 if the site information is consistent with the statement. Cumulate item scores across the scale to yield a total evidence-based score for the site.

2. Rating Criteria Used in This Study

The following rating criteria were modified based on the criteria used in (Griffiths & Christensen, 2005). They are a common standard for both human raters and automated rating approaches to conduct quality score rating. As explained in section 3.4, some evidence-based treatment guidelines were split into multiple criteria since the original guideline item contain multiple semantic propositions.

In addition, a general rating guideline is that the matching between a sentence and a criterion is based on checking the coverage of the main point of a criterion. The matching of modifiers in the criteria is nice to have, but not mandatory. In the following rating criteria, the modifiers are the parts included by brackets.

1. Antidepressant medication is an effective treatment for major depressive disorder.
2. Antidepressants are all equally effective.
3. The effectiveness of antidepressants is around 50 to 60%.
4. Full psychosocial recovery can take several months.
5. Drop out rate is same for different antidepressants.
6. The side effect profile varies for different antidepressants.
7. The choice of antidepressant should depend on individual patient factors (e.g. presence of co-morbid psychiatric or medical conditions, previous response to a particular drug, patient preference regarding the desirability of specific side-effects, concurrent drug therapy, suicidal risk)
8. Antidepressants are not addictive.
9. A trial of 6 weeks (at full dose) is needed before an antidepressant can be considered to have failed, or another antidepressant can be considered to try. [Score 1 if mention at least one of these]
10. A second-line drug should probably be from a different class of antidepressant.
11. Once improved continuation treatment (at the same dose for at least 4-6 months) should be considered.

12-A. Antidepressants should not be stopped suddenly.

12-B. Abrupt cessation can cause discontinuation syndrome.

13-A. St John's Wort appears to be as effective as (tricyclic) antidepressants.

13-B. St John's Wort causes fewer side effects than (tricyclic) antidepressants.

13-C. Little is known about any long term adverse effects of St John's Wort.

14. Cognitive therapy can be an effective treatment for depression.
15. Cognitive behaviour therapy is (at least) as effective as antidepressant treatment in (mild-to-moderate) depression.
16. Cognitive behaviour therapy may be valuable for people who respond to the concept of Cognitive behaviour therapy, prefer psychological to antidepressant treatment, have not responded to antidepressant therapy. [Score 1 if mention at least one of these]
17. Problem-solving may be effective for depression.
18. (Generic) counselling is probably no more effective than treatment as usual from the GP for depression.
19. Written information (usually based on a cognitive model of depression) can improve mild-to-moderate depression.
20. Exercise can be effective for depression (- alone or as an adjunct to other treatments).

Appendix D: Human Rating Code and Instructions

To guarantee high inter-rater and intra-rater reliability of rating results, the following rating code was to provide raters a guide to help them comply with the same set of decision patterns while matching sentences with evidence-based criteria.

Rating Codes for Raters to Follow:

1. The general task is to read the text and to identify if the idea of any rating criteria is contained in the text.
2. The unit of analysis is sentence. Matching of sentence meaning against rating criteria needs to be completed by examining whether the main idea of a rating criterion is presented by a sentence.
3. Modifiers in rating criteria which are not essential parts affecting the main idea are enclosed by bracket (available in Appendix C). The missing of these modifiers in sentence should not affect raters' decision on criteria matching.
4. Raters need to comply with the following code when matching between general concepts and specific instances ---
    a. It is valid to project the general concept (e.g. antidepressant) in the rating criteria to specific instances (e.g. tricyclic antidepressant) in the text because if a proposition for general is TRUE, it will also be TRUE by replacing the general with specific instance.
    b. It is invalid to project the specific instance (e.g. St. John's Wort) in the rating criteria to general concepts (e.g. herbal) in the text simply because a TRUE proposition for specific instance may not be TRUE for sibling instances.
5. Raters need to avoid relying on logic inference to do meaning matching since logic inference may extend meaning to a scope which is not necessarily presented by the original text.

    For example, text

    > "Age, sex, body size, body chemistry, physical illnesses and their treatments, diet, and habits such as smoking, are some of the factors that can influence a medication's effect."

    Criteria #7

    > "the choice of antidepressant should depend on individual patient factors"

    They do not exactly talk about the same issue. Raters should not equal the text to the criteria by making logical inference based on their personal assumptions.

6. Raters need to label the sentence or sentence group of which they believe the meaning is consistent with the rating criteria. A pair of tags was used to enclose the identified sentences. For example, <Criterion 1> *sentence content* … </Criterion 1>
7. If the rater believes that it is a group of continuous sentences, instead of a single sentence alone, that conveys the meaning of a rating criterion, then label the sentence group.
8. Within the labeled sentences, raters need to underline the key words or phrases which raters regard as essential semantic elements for matching with criteria.
9. Quality scoring - score 1 for each criteria. Scoring policy for individual criterion is specified in rating criteria (available in Appendix C).
10. For each rated page, raters need to summarize the number of matching criteria contained in the page.
11. Raters need to complete rating independently.
12. After independent rating completed, two human raters need to conduct cross-review on rating results to identify discrepancies.
13. Raters meet to discuss discrepancies and to achieve agreement. When an agreement could not be reached, researcher reviewed data and then made final decision.

Curriculum Vitae

Name:	Yanjun Zhang
Post-secondary Education and Degrees:	Nanjing University of Aeronautics and Astronautics Nanjing, Jiangsu, China 1992-1996 BEng, Electrical Engineering Dalhousie University Halifax, Nova Scotia, Canada 2002-2004 M.Sc Computer Science The University of Western Ontario London, Ontario, Canada 2004-present Ph.D. Library & Information Science
Honours and Awards:	Social Science and Humanities Research Council (SSHRC) 2006-2007, 2007-2008
Related Work Experience	Teaching Assistant The University of Western Ontario 2004-2005, 2005-2006 Research Assistant The University of Western Ontario 2006-2007

Publications:

Vaughan, L. & Zhang, Y. (2007). Equal representation by search engines? A comparison of websites across countries and domains. *Journal of Computer-Mediated Communication, 12(3),* 888-909.

Zhang, Y. (2006). The effect of open access on citation impact: A comparison study based on web citation analysis. *Libri*, *33(3)*, 145-156.

Jutla, D.N., Bodorik, P. & Zhang, Y. (2006). PeCAN: An architecture for privacy-aware electronic commerce user contexts, *Information Systems*, Elsevier, Special issue on the Semantic Web and Web Services, *31(4)*, 295-320.

Jutla, D.N. & Zhang, Y. (2005). Maturing e-privacy with P3P and context agents, *IEEE International Conference on E-Technology, E-Commerce and e-Service (EEE 2005)*, *Hong Kong, March 29-April 1, 2005* (pp. 536-541).