HealthTrust: Assessing the Trustworthiness of Healthcare Information on the Internet

By

Meeyoung Park

Submitted to the graduate degree program in Electrical Engineering and Computer Science
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Committee members

____
Bo Luo, Chairperson

____
Xue-wen Chen

____
Arvin Agah

____
Luke Huan

____
Prasad Kulkarni

Date defended: ______________________________

The Dissertation Committee for Meeyoung Park certifies
that this is the approved version of the following dissertation :

HealthTrust: Assessing the Trustworthiness of
Healthcare Information on the Internet

____

Bo Luo, Chairperson

Date approved: ____________________

Abstract

As well recognized, healthcare information is growing exponentially and is made more available to public. Frequent users such as medical professionals and patients are highly dependent on the web sources to get the appropriate information promptly. However, the trustworthiness of the information on the web is always questionable due to the fast and augmentative properties of the Internet. Most search engines provide relevant pages to given keywords, but the results might contain some unreliable or biased information. Consequently, a significant challenge associated with the information explosion is to ensure effective use of information. One way to improve the search results is by accurately identifying more trustworthy data. Surprisingly, although trustworthiness of sources is essential for a great number of daily users, not much work has been done for healthcare information sources by far.

In this dissertation, I am proposing a new system named HealthTrust, which automatically assesses the trustworthiness of healthcare information over the Internet. In the first phase, an unsupervised clustering using graph topology, on our collection of data is employed. The goal is to identify a relatively larger and reliable set of trusted websites as a seed set without much human efforts. After that, a new ranking algorithm for structure-based assessment is adopted. The basic hypothesis is that trustworthy pages are more likely to link to trustworthy pages. In this way, the original set of positive and negative seeds will propagate over the Web graph. With the credibility-based discriminators, the global scoring is biased towards trusted websites and away from untrusted websites. Next, in the second phase, the content consistency between general healthcare-related webpages and trusted sites is evaluated using information retrieval techniques to evaluate the content-semantics of the webpage with respect to the medical topics. In addition, graph modeling is employed to generate contents-based ranking for each page based on the sentences in the seed pages. Finally, in order to integrate the two components, an iterative approach that integrates the credibility assessments from structure-based and content-based methods to give a final verdict - a HealthTrust score for each webpage is exploited. I demonstrated the first attempt to integrate structure-based and content-based approaches to automatically evaluate the credibility of online healthcare information through HealthTrust and make fundamental contributions to both information retrieval and healthcare informatics communities.

Acknowledgements

I would like to express my gratitude to these faculties, colleagues and my families for their support and assistance with my dissertation.

First of all, I would sincerely like to thank my advisor Dr. Bo Luo for his valuable guidance and encouragement during my graduate studies. He was a great mentor, and very supportive while guiding me. He discussed with me enthusiastically and emphasized on critical thinking to solve problems. Also I'd like to thank to my co-advisor, Dr. Xue-wen Chen. When I started my Ph.D at the University of Kansas, he put his effort for me to serve as a Graduate Teaching Assistant and provided me with financial support for my conference travel. I learned a lot from him, not only research skills but also his dedication towards research. I also gratefully thank my dissertation committee members: Dr. Arvin Argh, Dr. Luke Huan, Dr. Prasad Kulkarni and Dr. Michael Wang. They offered professional and valuable suggestions on my proposal and dissertation.

Second, I would also like to thank all my colleagues in our research group, especially Lei Yang, Yuhao Yang, Jongcheol Jung and Hariprasad Sampathkumard. These friends and co-workers helped a lot with my research. We were working together in the lab very late and discussed on our research topic actively with each other. Also, they were willingly helped me out whenever I needed help.

Third, I want to thank Dr. Chungoo Park who is an assistant professor in Chonnam National University. He always encouraged me whenever I had a hard time during my gradate studies and helped me to dig deeper in my discipline through our discussion on various Bioinformatics topics. I appreciate Dr. Eva Feldman who is a professor in Neurology Department at the University of Michigan to offer me a postdoctoral research fellow position. Without her help, I could not have a fulfilling and successful defense of my dissertation.

Finally but the most importantly, I want to thank my parents Kyungsuk Park, Nasoon Lee and two elder sisters and two younger brothers. My parents gave me all the encouragement and endless love to support my education from elementary school to college. Their unselfish love makes me stronger and warmer even I studied far away from them. My husband, Jongwook Kim, inspired me to have ambition and provided me with endless encouragement and financial support in my research and study.

Contents

List of Figures

List of Tables

Chapter 1

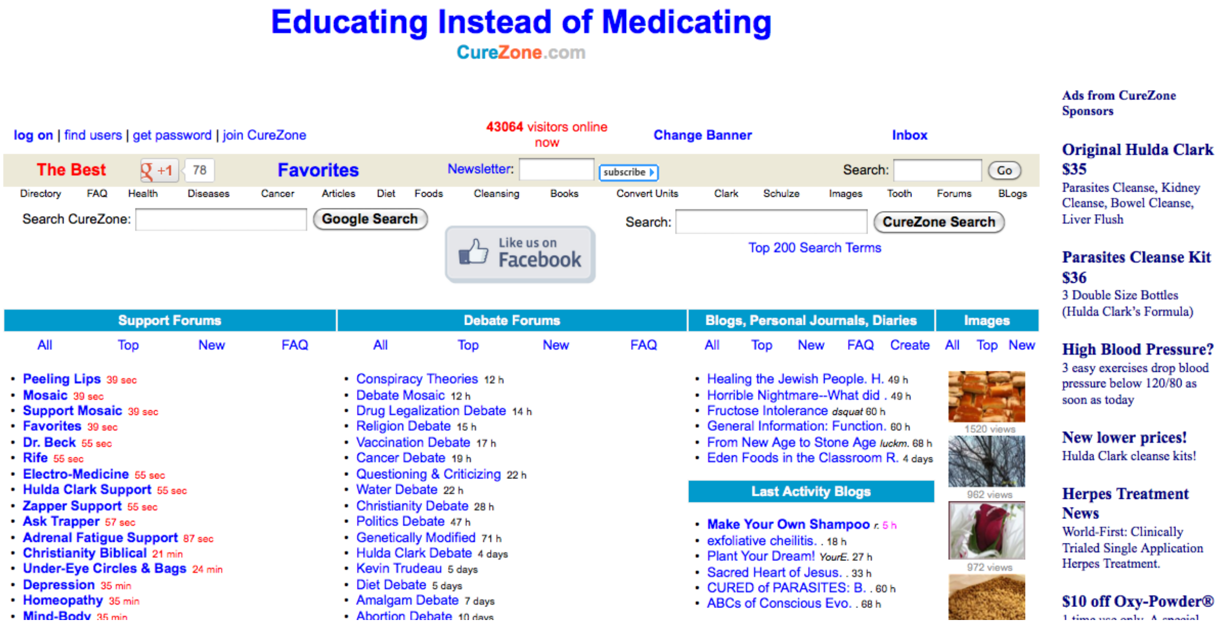
Introduction

The rapid growth of Web 2.0 and social networks has changed the way people seek health information. Instead of relying on traditional media such as TV, radio and newspaper, users now satisfy their information needs through search engines, social media (blogs, tweets, etc.) and wikis. More than 70,000 Websites provided health information and more than 50 million people searched health information on the Internet in 2001 [10], and these numbers have been increased as of now. In this context, huge amount of healthcare related information has been published through various sources: government agencies, non-profit organizations, hospitals, clinics, pharmaceutical and insurance companies, and producers of other health related products. Alternatively, there are huge volumes of personal websites, blogs and tweets that introduce personal experiences and advices from patients, doctors, nurses, and product sales. As Robinson et al. [51] defined, *seeking* health information is an interactive communication between consumers, patients, professionals and computers or other mobile devices such as smart phones. Empirical studies were conducted to assess the quality of the information on the Web and found out the dependency of patients or consumers on the Internet as their medical references [50, 17, 24, 23]. The results showed that the dependency of the websites was highly related to the websites' design or maintenance style rather than the quality of contents.

1.1 Motivation

Even though in Healthcare, scientific, accurate and objective information is vital, not all health-related content on the Internet is trustworthy. Healthcare-related government sites such as NIH publish information that is highly scientific, but sometimes they could be difficult for average users to understand. Non-profit organizations maintain websites with inconsistent quality control. Hospitals, clinics, pharmaceutical and insurance companies hold websites mostly for their business purposes, but these sites also contain general healthcare information that might be somewhat biased. Producers of other health related products (e.g. herbal or dietary supplements) may aggressively advertise their products and publish information that is exaggerated in favor of their own interests. For example, Figure 1.1 apparently looks a trustworthy website containing all kinds of medical information which is found using "heart attack" on Google search engine. However, it is most likely that one can determine the suspiciousness of the site since they use very vague and suspicious words such as "cure within 30 days" and contain too many advertisements in the website. Also they provide various forums and debates but, healthcare-related blogs or forums are often filled with information of irregular quality. In particular, with the extreme popularity of Web 2.0 and social networking, Websites such as Facebook, Myspace and Twitter have become very influential in users information *seeking* behaviors, in some cases becoming the primary information source for many users. However, healthcare-related information in social networking sites is highly inconsistent in quality.

Even though credible websites like the Center for Disease Control and Prevention (CDC) or the National Institute of Health (NIH) guarantee the trustworthiness of healthcare information they provide, their use of medical terminology can sometimes be hard for users to understand. Users may tend to prefer more explanatory sites such as Wikipedia, a free editable online encyclopedia. Furthermore, Google often returns Wikipedia pages as the first page in its search results. Recently, a small-scale study conducted by Leithner et al. examined the quality of the Wikipedia articles relating to Osteosarcoma [37] in comparison



- Peeling Lips 39 sec
- Mosaic 39 sec
- Support Mosaic 39 sec
- Favorites 39 sec
- Dr. Beck 55 sec
- Rife 55 sec
- Electro-Medicine 55 sec
- Hulda Clark Support 55 sec
- Zapper Support 55 sec
- Ask Trapper 57 sec
- Adrenal Fatigue Support 87 sec
- Christianity Biblical 21 min
- Under-Eye Circles & Bags 24 min
- Depression 35 min
- Homeopathy 35 min
- Mind-Body 35 min

- Conspiracy Theories 12 h
- Debate Mosaic 12 h
- Drug Legalization Debate 14 h
- Religion Debate 15 h
- Vaccination Debate 17 h
- Cancer Debate 19 h
- Questioning & Criticizing 22 h
- Water Debate 22 h
- Christianity Debate 28 h
- Politics Debate 47 h
- Genetically Modified 71 h
- Hulda Clark Debate 4 days
- Kevin Trudeau 5 days
- Diet Debate 5 days
- Amalgam Debate 7 days
- Abortion Debate 10 days

- Healing the Jewish People. H. 49 h
- Horrible Nightmare--What did . 49 h
- Fructose Intolerance *dsquat* 60 h
- General Information: Function. 60 h
- From New Age to Stone Age *luckm.* 68 h
- Eden Foods in the Classroom R. 4 days

- Make Your Own Shampoo *r.* 5 h
- exfoliative cheilitis. . 18 h
- Plant Your Dream! *YourE.* 27 h
- Sacred Heart of Jesus. . 33 h
- CURED of PARASITES: B. . 60 h
- ABCs of Conscious Evo. . 68 h

Ads from CureZone Sponsors

Original Hulda Clark $35
Parasites Cleanse, Kidney Cleanse, Bowel Cleanse, Liver Flush

Parasites Cleanse Kit $36
3 Double Size Bottles (Hulda Clark's Formula)

High Blood Pressure?
3 easy exercises drop blood pressure below 120/80 as soon as today

New lower prices!
Hulda Clark cleanse kits!

Herpes Treatment News
World-First: Clinically Trialed Single Application Herpes Treatment.

$10 off Oxy-Powder®
1 time use only. A special

Figure 1.1: An example of vulnerable website

to the ones available in National Cancer Institute (NCI). They observed the quality of the Wikipedia articles to be good and more accessible than the NCI articles, but found them to lack scientific citations.

The most popular step of seeking health information for consumers is using search engines. The important roles of search engines in this field were discussed in [46, 31, 54]. Meric et al. [40] used the key words, "breast cancer" on Google and examined the first 200 websites over 100,00 English sites. They evaluated the characteristics of the websites and showed that only 57% had the authorship in the web pages and rest of them showed partially or none. In addition, the study confirmed that the quality of the information had no correlation with link popularity.

We also evaluated with three different kinds of keywords using Google. Firstly, when we put unusual search terms such as "sprain field treatment", "squirrel bite", out of 13 top pages, eight are forum sites or answers from portal sites, which is mostly 'believe or not', one authoritative, and four unrelated sites. Secondly, we tried sentences, for example, "I want to know about Lasik surgery", eight out of top 13 pages are from commercials to connect specific eye doctors, there are forum or blogs and one is "unrelated", rest of them are "news". The unrelated site has no information about the Lasik surgery. Also, we used general terms like "flu treatment", "health care medicine", and "heart attack medicine". The results shown with top priorities are from insurance companies, job search, drug websites that direct you to many other commercial sites. However, with the developments of information retrieval technologies, search engines are well improved to design to assess the relevance as well as the importance (or authority/popularity) of web pages. But, they are not currently using credibility as a factor in ranking.

Apart from the quality of health information available in search engine results and Wikipedia, we also need to consider the quality of information available in social networks. Since more and more people are engaged in their use, their content also plays an important role in the dissemination of health information among general consumers. Weitzman et al. conducted a study to observe the quality and safety of diabetes-related social networks. [60] They reported the quality to be variable, but found security and privacy of user's personal data to be poor. Although the study was conducted on a small scale, it is enough to show that social networks inevitably contain suspicious healthcare information. In order to address these issues, we need new automated approaches for a scientific and objective measurement of trustworthiness of healthcare information.

In this way, uninformed users become very vulnerable when they search for health information on the Internet. Therefore, the quality of online healthcare information becomes a concern due to the lack of quality control on the web. It has been observed that a very large portion (more than half) of online healthcare information sources provide inaccurate information [3]. For example, from a variety of types of unreviewed sources, some information materials are provided by authors without professional training [13], some adopt "a patronizing tone to promote a participative approach to decision making" [22, 11], and most of the others are not reliable due to "lack of context" [21]. While an increasing number of critics question the quality of online health information, limited insights has been provided.

However, there have been alternative ways to protect health information consumers. For Instance, policy makers and government agencies (e.g. FDA) have made efforts to prevent producers and retailers from distributing exaggerated or inaccurate information on healthcare-related products. Medical Library Association [2], a non-profit organization website, provides a guide to evaluate the health information on some popular websites on the Internet [8], but all sites could not be evaluated as needed. In addition, due to the excessive amount of information available on the Internet, it is extremely difficult to implement effective surveillance mechanisms to enforce such policies. Meanwhile, they are unable to regulate personal opinions posted on blogs and forums. On the other hand, social voting has been very successful in many applications, e.g., retailer and product ratings, social recommendations, however, it is not suitable for judging the credibility of online healthcare information.

With the acknowledgement of the problems in seeking health information, it is highly expected to have a mechanism that automatically rates the trustworthiness of healthcare information over the Internet. Consider that some search engines give a warning on suspicious (spam or virus) Websites; likewise, it is desired if it can be delivered such an assessment of credibility for healthcare-related contents. In order to design effective approaches to assess the credibility and trustworthiness of online healthcare information, as well as to help people recognize and utilize such information, firstly, a thorough understanding of healthcare information widely distributed over the Web is needed, in particular, a study of the providers of such information. The study includes who they are, how they are distributed, how they are related (i.e. how they cite (endorse) their peers), etc. Graph analysis has been employed to study various types of data, such as the Internet, online social networks, etc. In this study, we collect online healthcare-related information through a focused crawler, and apply statistical, graph, and link analysis methods to explore and analyze such data. Last but not least, we design two case studies that ask users to evaluate search results from commercial search engines. This preliminary work is addressed in detail in Chapter 3.

Furthermore, Natural Language Processing (NLP) and Machine Learning (ML) are now essential techniques to process text data in medical informatics. [52, 53, 28] In this study, we propose a content-based analysis using the above techniques to analyze healthcare text data on the web and assess its credibility. Our proposed method is inspired by an observation: websites whose content are similar to trusted websites are also more likely to be trustworthy. Therefore, our approach tries to identify similarities in website content in comparison to content from known websites. To do this, we first gather healthcare related pages from the internet using a focused crawler. We then use two methods: HMM based sentence models to identify the trustworthiness of healthcare information and a "Bag-of-words" based Topic Discovery method to identify topics within the sentences of those pages. We then perform page-level and site-level classifications based on results from both these methods to identify the trustworthy and suspicious sites. We evaluated our method on randomly chosen real dataset and are able to achieve about 90% accuracy in identifying the trustworthiness of the content.

1.2 HealthTrust

In this dissertation, a new system named HealthTrust, which automatically assesses the trustworthiness of healthcare information over the Internet, is proposed. In Phase I, structure-based analysis is performed. In order to do that, firstly, unsupervised clustering using graph topology on our collection of healthcare-related websites is employed. The goal is to identify a relatively larger and reliable set of trusted websites without much human efforts. Our method starts with Affinity Propagation (AP) [27] clustering, which represents an individual data point as a node in the network, then uses belief propagation methods that recursively communicate with real-valued messages along edges until clusters emerge. It is expected (and proved in a small-scale preliminary work) that some highly trustworthy websites (e.g.

cdc.gov and fda.gov) are categorized into a few clusters, while some obviously spam pages are also clustered. Such clusters will be easily identifiable and used as (positive and negative) seeds in the future steps.

Based on the seed set, a biased TrustRank algorithm for link-based assessment is developed. The basic hypothesis is that trustworthy pages (originated from the seeds) are more likely to link to trustworthy pages. This phase starts with an existing approach, namely TrustRank [32], which essentially requires the teleport operation in PageRank to be destined for trusted seed pages only. Other approaches will be explored to impose the discriminator into the original PageRank algorithm, so that endorsements (in-links) from trustworthy pages will carry higher weight, while endorsing (out-link to) untrusted pages is a negative factor. In this way, the original set of positive and negative seeds will propagate over the hyper-link graph. With the credibility-based discriminators, the global scoring is biased towards trusted websites and away from untrusted websites.

Next, in Phase II, two novel approaches based on topic modeling and machine learning techniques have been emplyed to assess the trustworthiness of the information provided in healthcare sites by doing content-based analysis automatically. The preliminary study has shown that term distribution similarity will not generate satisfying results since trusted and suspicious pages use very similar terms to express opposed opinions. To tackle such problem, two analysis methods have been done: (1) *Topic discovery*: we make use of TAGME to identify salient topics in the sentences available in the healthcare websites. An analysis of the similarity measures among the topics identified is used to decide if the information from candidate website falls under the suspicious or trustworthy category.; and (2) *HMM analysis*: apply Hidden Markov Models to model trustworthy and suspicious sentences using an annotated training set.

Finally, in order to integrate the system, an iterative approach that integrates the credibility assessments from structure-based and content-based methods to give a final verdict - a HealthTrust score for each website will be exploited. In the iterations, strongly positive and strongly negative results from the structure-based approach will be used as "additional seeds" in the content-based approach, and vise versa. The iterative approach further counteracts the problem of limited seeds as well as the sparseness of the document space.

1.3 Contributions

HealthTrust aims to identify the trustworthiness of the healthcare related information on the Internet automatically and provide more credible guidance to consumers. Therefore, success of this work will bring great benefits to the general consumers. Especially, our contributions in the content-based analysis are primarily three fold: (1) We have proposed two novel approaches for performing content based analysis on healthcare data. (2) We have been able to show that the Topic Modeling approach is able to perform better than the HMM approach due to its ability to effectively capture semantic information and (3) the algorithm for performing content analysis scales linearly making it suitable for handling big data. The contributions of this research are below.

- We use a new algorithm to propagate trust in a two-way Web graph efficiently.
- We discover semantic topics for content analysis of healthcare information.
- We integrate the structure and content-based methods efficiently without loss of their orthogonal properties using a new iterative algorithm expanding the positive and negative seed sets automatically.
- We have proposed two novel approaches for performing content based analysis on healthcare data.
- We have been able to show that the Topic Modeling approach is able to perform better than the HMM approach due to its ability to effectively capture semantic information.
- The proposed algorithm for performing content analysis scales linearly making it suitable for handling big data.

The general contributions of the HealthTrust:

- Demonstrates the first attempt to integrate link-based and opinion-based approaches to automatically evaluate the credibility of online healthcare information through HealthTrust.

- Makes fundamental contributions to both information retrieval and healthcare informatics communities.

- Builds a first milestone in trusted public healthcare information management.

- Promotes trustworthy sources that provide creditable information, in addition to helping the general consumers to interpret healthcare-related information over the Internet.

The rest of the dissertation is organized as follows. First, an overview of the related work that motivated the development of the HealthTrust is given in Chapter 2. Chapter 3 provides the preliminary work for data collection and observation of the data. Chapter 4 shows how we select the seed sets for ranking the trustworthy sites. Chapter 5 provide the overall analysis for the healthcare information web graph using the existing methods. The details on our method is provided in Chapter 6. Chapter 7 shows experiment results after performing our method on real data set. Finally, Chapter 8 summarize what we have done, what our contributions are, and what can be done in the future.

Chapter 2

Related Work

In this section, we provide an overview of the two approaches that motivated the development of the HealthTrust system. The first approach is ‘link analysis’ inspired by the observation that websites cited by trusted websites (e.g., CDC, NIH) are more likely to be trustworthy, while websites citing spam sites are suspicious. The second is ‘Semantic Analysis’ inspired by the observation that websites whose opinion is consistent with trusted websites are more likely to be trustworthy. Followed the two approaches is a brief review of the Affinity Propagation clustering method adopted in our HealthTrust.

2.1 Link Analysis

Link analysis is the method that extracts knowledge from a network or a graph by analyzing its structure which is consisted of nodes and links. By doing so, in-depth insights are provided intuitively that help us identify the key components or objects within the network. The applications of link analysis are very diverse from natural sciences, such as biology, and pharmacology to modern technology or crime analysis, such as telecommunication network analysis, fraud detections of bank, or insurance company. In particular, the World Wide Web, or Web, is considered as a huge graph structure due to its nature of hyperlink function in Hyper Text Markup Language(HTML) [5] like <a href =“www.google.com”>. The

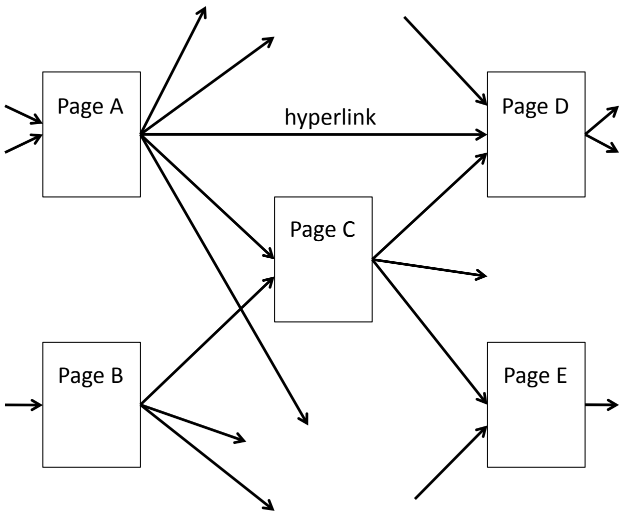


Figure 2.1: A graph of Web

hyperlinks allow web pages to link to or connect with each other. In information retrieval, to assess the importance or authority of webpages, link analysis is employed to study the link-based relationships between nodes. A link from node A to node B is often treated as an *endorsement* or vote to support B. Figure 2.1 shows an example of a directed graph of the Web. Each page contains one or more hyperlinks that point to other pages. The number of incoming links is the in-degree of the node and the number of outgoing links is the out-degree of the node. If there's no outgoing links in the page, the node is called 'dangling node' or 'terminal node'.

The analysis of the Web graph was motivated by scientific citation analysis [19, 20, 44], which is used for a measurement of citation ranking among scientific journals, such as *impact factor*. The measurement is solely dependent on counting the number of incoming links in the network within a specific time period. The citation analysis boosted up the development of the ranking algorithms in the Web. Ranking algorithms simply consider links of the web graph as citations of the academic literatures. Based on the number of links, the relative importance of the page in the web graph can be ranked. The most popular ranking algorithms are PageRank [45] and HITS(Hyperlink-Induced Topic Search) [33]. Many other algorithms related to a ranking problem have been proposed so far including [**?** 42, 18, 6]. However, the basic concepts of those algorithms are based on PageRank or HITS. Therefore, the focus of this proposal is on PageRank and HITS that are reviewed in this section in detail.

2.1.1 PageRank

PageRank [45, 7] measures the probability that a page will be visited by a"tireless web surfer". Let $G = (V; E)$ be a directed graph that consists of a set of N pages (V) and a set of directed links between pages (E). The transition matrix T is defined as: $T(u; v) = 1$ if there is an edge from page v to page u and $T(u; v) = 0$ otherwise. The PageRank vector R for each page is then computed by Equation 2.1.

$$\mathcal{R}(u) = \alpha \cdot \sum_{v:(v,u)\in E} \frac{1}{\omega(v)} \cdot \mathcal{R}(v) + (1-\alpha)\frac{1}{\mathcal{N}} \tag{2.1}$$

where α is the decay factor for teleporting probability and $\omega(v)$ is the number of outgoing links from v. In an iterative calculation, $R(u)$ will eventually converge to the PageRank of u. Figure 2.2 is a simplified example of PageRank calculation without teleporting factor. As seen in the Figure, the rank of each page is the summation of all the scores of the incoming links. Each incoming link score is calculated by the division of the number of outgoing links of the page's rank. Page C has the highest rank in the graph followed by Page D and Page E, indicating that they are the three most important. For more detailed review, refer to [36, 4].

In spite of its popularity, PageRank has its own drawback: its negligence of the trustworthiness of webpages, causing unavoidable biased or untruthful pages. For example, untrustworthy sites can intentionally manipulate hyperlinks that point to or from good pages. In addition, they could create many incoming links that point to another vulnerable sites to make them important sites. Eventually, those sites cannot be filtered by PageRank algorithm. To alleviate the limitation, Gyönyi *et al.* [32] proposed a biased PageRank algorithm, called TrustRank, to distinguish good pages and reduce spam pages in the searched results. TrustRank is reviewed in Section 2.1.3.

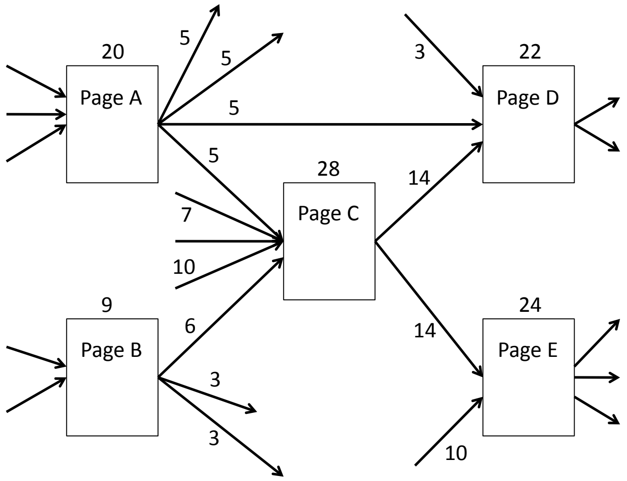


Figure 2.2: Simple Example of PageRank Calculation

2.1.2 HITS

HITS [33] is another popular link analysis approach which divides webpages into "hubs" and "authorities". Different from PageRank, it assumes that authoritative pages do not necessarily point to other authoritative pages. So they define a special node, *hub*, as "the page that has links to multiple relevant authoritative pages". By doing so, it can capture the global nature of Web finding central pages. Thus, it is suitable for a broader search rather than exact key word query search but also getting irrelevant. In many cases, it is possible for one to use a specific query to search for relevant pages containing the exact words. In HITS structure, each page can be both a hub and a authority, and the hub acts as a pointer to meaningful pages and authority shows the meaning of the page itself. Based on the hub-authority relationship, HITS can filter the unrelated page having large incoming links. It implements the idea that "good hubs" point to "good authorities", while "good authorities" are pointed to by good hubs. Hub and authority scores are thus calculated through an iterative approach.

Basically, given the query, all pages containing the query is gathered, and the pages are called a *root set*. Next, the root set is expanded to include any pages that point to a page in the root set and are pointed by a page in the root set. The expanded set is called a *Base set*.

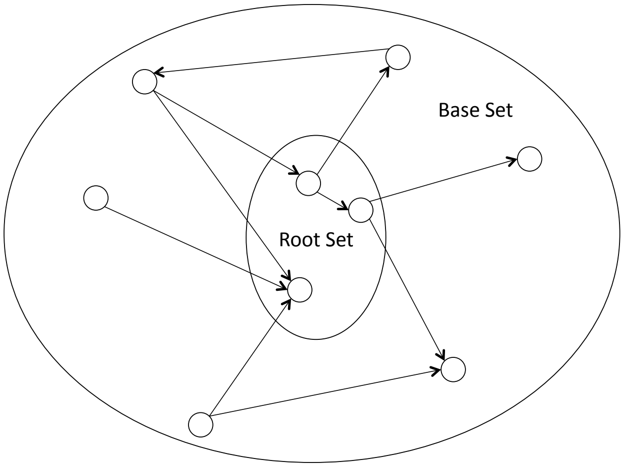


Figure 2.3: The expansion of the root set to a base set

Figure 2.3 shows the root set expansion to get a base set. The computation is iteratively performed for a hub score, $h(x)$, and an authority score, $a(x)$ for each page in the base set using Equation 2.2 and Equation 2.3 respectively. The hub score, $h(x)$ is calculated by summing up the authorities of the nodes that are pointed to by the hub, and the authority score, $a(x)$ is the sum of the hubs that point to this authorities. We implemented HITS algorithm, and applied it on the collected of healthcare information network. The results are shown in Chapter 5.

$$h(j) = \sum_{i \in B(j)} a_j \tag{2.2}$$

$$a(i) = \sum_{j \in B(i)} h_j \tag{2.3}$$

2.1.3 TrustRank

TrustRank is a semi-automatic ranking algorithm considering the trustworthiness of webpages proposed by Gyönyi *et al.* [32]. The basic hypothesis of TrustRank is that mostly good pages are likely to point to good pages. Figure **??** is an example of the good and bad sites relationship. Initially, they use manually curated seed set by human experts. The seed set

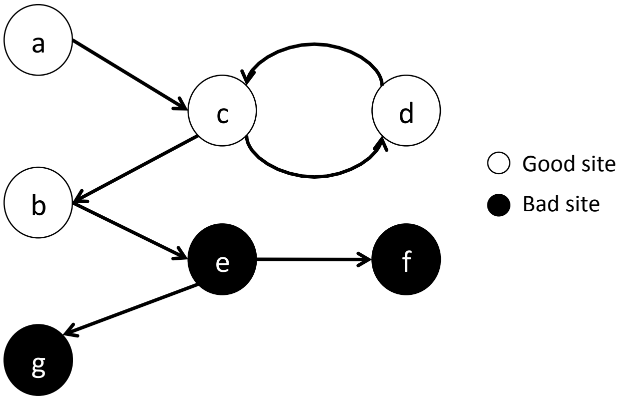


Figure 2.4: An example of a graph with good and bad sites

is chosen from the top-ranked pages produced by the inverse PageRank which is generated using the transition matrix of the PageRank. As the name showed, the inverse PageRank is inverted the link directions from the original link structure. The rational of the inverse PageRank is that the more out-links the page has, the more trustworthy. Then, for next step, TrustRank calculates the trust score using Equation 2.4. TrustRank defines the biased factor **d** to implement the dampening and splitting of trust in the graph in place of the decay factor α in PageRank. It means that the trust of a certain page will reduce if it is far away from the good seed pages. Vector **d** is defined as $d_i = 1$ if page i is selected as a good page and $d_i = 0$ if not. Then **d** is normalized by $|d|$. By doing so iteratively, it propagates the trust scores over the web graph.

$$\mathcal{R}(u) = \alpha \cdot \sum_{v:(v,u)\in E} \frac{1}{\omega(v)} \cdot \mathcal{R}(v) + (1-\alpha)\mathbf{d} \tag{2.4}$$

TrustRank has several limitations; firstly human experts must involve in deciding whether a page is good or bad and the decision might be biased and costly. Secondly, TrustRank uses the inverse PageRank to select desirable pages as seed sets. However, the inverse PageRank uses the out-links of the original graph that inevitably includes bad sites because some spam sites can deceive out-links to mislead a search engine ranking system.

Unfortunately, the ranking algorithms for the search engines so far analyze the link structure instead of the contents of websites. As a result, consumers looking for critical health information might get unwanted pages in the first several result pages. A worse scenario is that the top search results are mostly unrelated sponsor's websites, or YouTube videos. In order to overcome the current shortcoming of the ranking algorithms, the contents of the webpages should be analyzed to provide the consumers the degree of the trustworthiness of the page. In this proposal, therefore, opinion-based approach is adopted along with the link-based approach to understand the semantics of the web contents. However, before discussing our combined approach, we give a brief discussion of the opinion-based approach in the next section.

2.2 Affinity Propagation Clustering

Affinity Propagation clustering is proposed by Frey *et. al.* [27] which is a powerful unsupervised machine learning method for finding an optimal set of clusters using a new concept called *exemplar*. An exemplar is defined as "a data point that is nicely representative of itself and other data points". Basically AP algorithm considers an individual data point as a potential exemplar in the cluster. AP algorithm performs iteratively until it detects good exemplars efficiently and rapidly by exchanging messages between nodes from the network. Furthermore, instead of using common similarity distance such as Euclidian distance [15], users can define any pair-wise similarity measures. As the most distinct approach compared to existing clustering methods such as k-means clustering [38], AP does not require the initial selection of centers randomly, which might lead to a potential failure of clustering. Instead it uses actual data points as potential centers and uses belief propagation [63] methods that recursively communicate with real-valued messages until clusters are found. It is particularly suitable for very large and sparse data. Since we want to find the representatives or most influential domains in the web graph as our seed set, AP is an appropriate method to achieve our goal.

We briefly explain the algorithm of the AP clustering here. First of all, the algorithm constructs a similarity matrix to measure the affinities between nodes. The similarity between two data points, say $S(A, B)$, shows how well the node B represents node A [41]. The optimal exemplars are chosen by the Equation 2.5. The net similarity $\mathcal{S}(\mathbf{c})$ is calculated by summing up all similarities of data points to its exemplar c and is maximized to identify the optimal exemplars.

$$\mathcal{S}(\mathbf{c}) = \sum_{i=1}^{N} s(i, c_i) + \mathbf{\Delta} \tag{2.5}$$

where $\mathbf{\Delta}$ is a dampen function to avoid oscillations of the algorithm. However, instead of using initial number of clusters, AP assigns a priori knowledge, P, that is the *preference* value for each node showing the goodness of the node as an exemplar. The preference P can be used as a control parameter; if P is big, it is likely to find more exemplars. Figure 2.5 is a face clustering example using AP clustering by Frey *et. al.*

2.3 Sentence Modeling

Sentence modeling is a challenging problem in Natural Language Processing. NLP is a broad area dealing with human-computer interaction problems using machine learning (ML), statistical inference, information retrieval (IR), automatic summarization, part-of-speech(POS) tagging, sentiment analysis, topic modeling and so on. Recently, NLP is being actively adopted in many medical research as well as healthcare informatics area [53, 28]. It is natural that the majority of data format of NLP is written text. Therefore, parsing is the important process of the text information to understand human languages as input data. In particular, since parsing can be done without complete understanding of language, it is prerequisite procedure for most NLP. One of the parsing methods is called 'part-of-speech tagging' which puts a label to each word with appropriate part of speech in a sentence such as 'noun', 'verb' and 'adjective'.

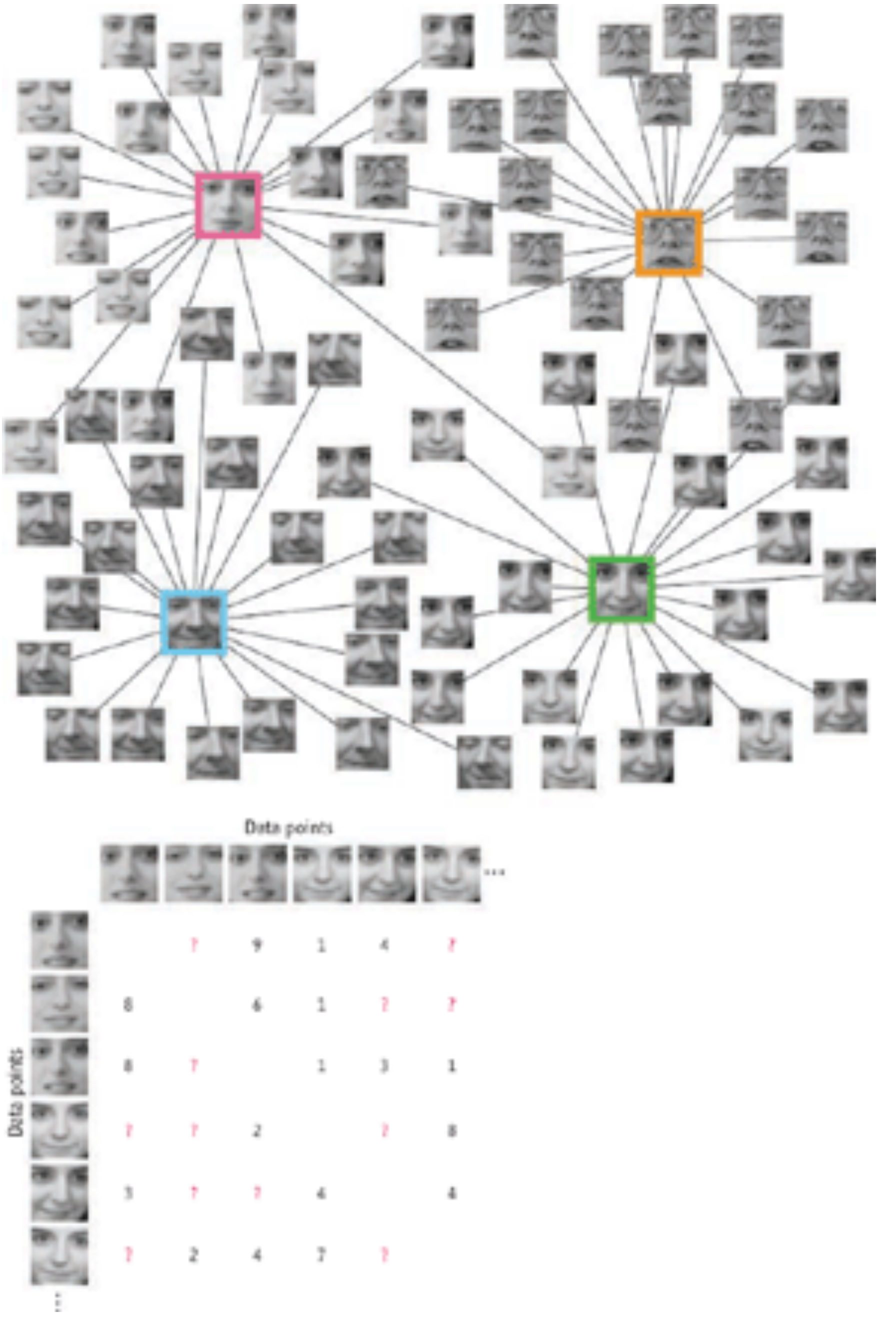
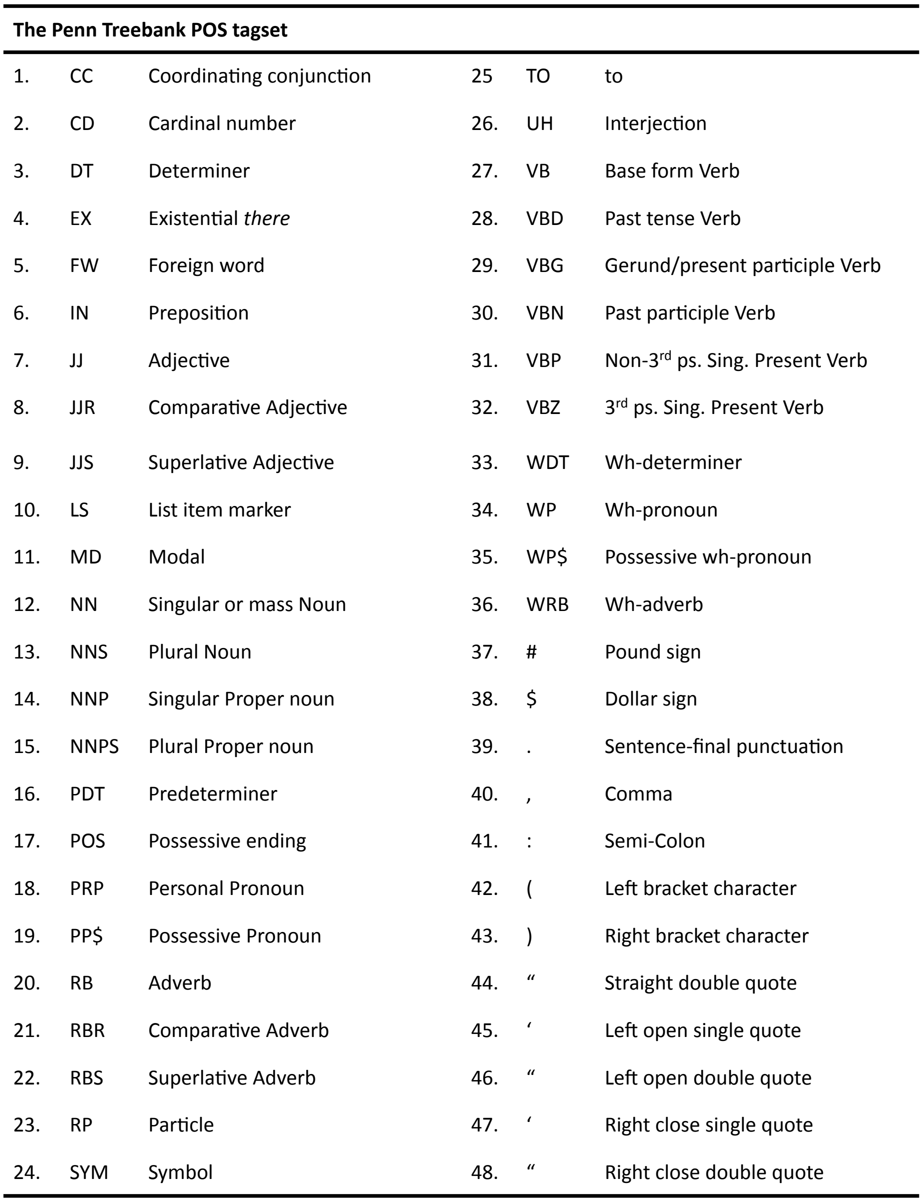


Figure 2.5: Face clustering [27]

2.3.1 POS Tagging

POS tagging is the process of tagging in English sentences. Originally, linguists made words of a language into several classes syntactically and labeled them as ‘nouns’, ‘verb’ and ‘adjectives’, and they are considered as parts of speech. Parsing techniques were pioneered from the two big corpus projects, Brown corpus [26] and the Penn Treebank project [39]. The Penn Treebank corpus contains over 4.5 million words of American English and is widely used as a reference tagging. Table 2.6 shows the tag set from the Penn Treebank corpus. Due to its huge size of data in Penn Treebank project, there were two steps in tagging all corpus; The first step was an automatic method that uses computer algorithm, and the second step required human annotators to correct the automatic task since language naturally contains ambiguity and inaccuracy. However, it is really laborious for human to do the correction process for tagging. In order to avoid the manual tasks, many ML techniques are adopted such as Hidden Markov Model (HMM) [35], Support Vector Machines (SVMs) [30], or Maximum-Entropy classifier [49, 58]. The detailed discussion of the methods is out of scope for this proposal, however, recent taggers give good results over 97% [57].

Although POS-tagging is the required process of analyze the textual information, it has fundamental drawbacks. First, POS tagging is syntactic analysis, hence, it is not enough to understand the meaning of the context. It causes ambiguity due to the complex nature of a language. An example of ambiguous cases in POS tagging is given in Figure 2.7. In the example, in the first figure, *training* is used as the main verb, and in the second and third figures, the main verb is *is*. To reduce the ambiguity, further analysis must be followed for understanding the text clearly as the context. Secondly, it produces same tagging results even though the text has different semantics. For example, the two sentences; ‘the supplement works good’ and ‘the supplement works bad’ have the opposite meaning, but the POS tagger gives the same results, such as NP-V-ADJ. The problem is due to two different adjectives, ‘good’ and ‘bad’. Therefore, after POS tagging, semantic analysis is inevitable. In the next section, we review *sentiment analysis* as one of the semantic analysis methods.



The Penn Treebank POS tagset					
1.	CC	Coordinating conjunction	25	TO	to
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Base form Verb
4.	EX	Existential *there*	28.	VBD	Past tense Verb
5.	FW	Foreign word	29.	VBG	Gerund/present participle Verb
6.	IN	Preposition	30.	VBN	Past participle Verb
7.	JJ	Adjective	31.	VBP	Non-3rd ps. Sing. Present Verb
8.	JJR	Comparative Adjective	32.	VBZ	3rd ps. Sing. Present Verb
9.	JJS	Superlative Adjective	33.	WDT	Wh-determiner
10.	LS	List item marker	34.	WP	Wh-pronoun
11.	MD	Modal	35.	WP$	Possessive wh-pronoun
12.	NN	Singular or mass Noun	36.	WRB	Wh-adverb
13.	NNS	Plural Noun	37.	#	Pound sign
14.	NNP	Singular Proper noun	38.	$	Dollar sign
15.	NNPS	Plural Proper noun	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Semi-Colon
18.	PRP	Personal Pronoun	42.	(	Left bracket character
19.	PP$	Possessive Pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	“	Straight double quote
21.	RBR	Comparative Adverb	45.	‘	Left open single quote
22.	RBS	Superlative Adverb	46.	“	Left open double quote
23.	RP	Particle	47.	‘	Right close single quote
24.	SYM	Symbol	48.	“	Right close double quote

Figure 2.6: The Penn Treebank POS tagset

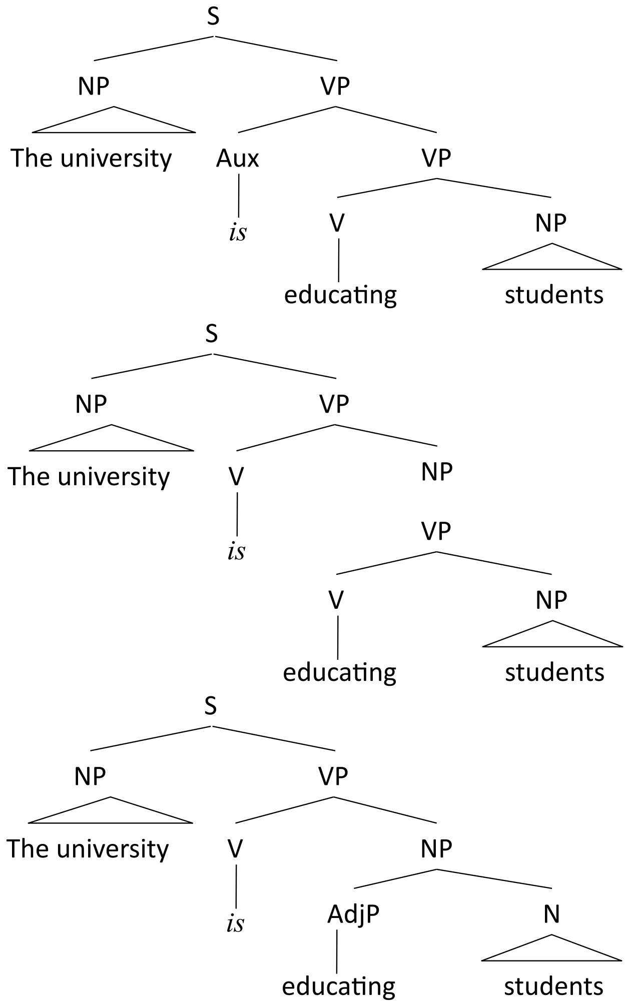


Figure 2.7: Ambiguous POS tagging

2.3.2 Hidden Markov Model Analysis

We used two approaches for our content-based analysis on healthcare data; Hidden Markov Model [48] and TAGME. [25]

2.3.3 Hidden Markov Model

A Hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with hidden states. Primarily, HMMs have been used to model sequence data like speech utterances in speech recognition. [47] They have also been used in Part-of-Speech tagging [14] and Named Entity Recognition [64] tasks. The success of HMMs in identifying patterns in sequential data has motivated us to explore the possibility of using HMM for content-based analysis. In general, a HMM can be defined using the following parameters:

Notation and definition

N: Number of states in the HMM

M: Number of observation symbols in the HMM

$A = [a_{ij}]$: N by N state transition probability matrix

$B = b_j\,(m)$: N by M observation probability matrix

$\Pi = [\pi_i]$: N by 1 initial state probability vector

An HMM is used to model a sequence with hidden states that represent the latent characteristics of the pattern that we are trying to model, which however emit symbols or observations that are visible. The outputs of the hidden states are observable and are represented as probabilistic functions of the state. In case of sentence modeling, the hidden states would represent the characteristics of a sentence, while the words forming the sentence would represent the visible observations. HMM is a supervised learning method where a training set is used to train the model. The Baum Welch algorithm is used for this and it learns the transition and observation probabilities of the HMM. Once trained, the HMM can then be used for computing the probability of a sentence belonging to given model using the Forward-Backward algorithm or can be used to predict the possible hidden state sequence that could have generated a given sequence of observations using the Viterbi algorithm.

2.4 Topic Modeling

2.4.1 Short-Text Tagging

Traditional topic modeling methods use probabilistic approaches based on "bag-of-words" model. [55] However, the "bag-of-words" approach is solely based on the frequency of terms in a document; therefore it is hard to capture the semantics of the text. In order to overcome the problem, Latent Semantic Analysis (LSA) [16], Explicit Semantic Analysis (ESA) [29] or Knowledgebase approaches [62] have been proposed. Recently, with the rapid growth of Wikipedia's knowledgebase and its link structure connecting the related concepts efficiently, several ESA based approaches using Wikipedia have been studied. [59, 56, 12] One of the ESA methods is TAGME, which is a web application tool for identifying underlying topics in short text fragments using Wikipedia and its link structure proposed by Ferragina and Scaiella (http://tagme.di.unipi.it/) [25]. They improved their method based on the studies of Kulkarni et al. and Cucerzan to deal with annotating very short texts or fragments such as tweets or news feed items *on-the-fly*. [34, 12]

2.4.1.1 TAGME

A systematic way of topic identification is using *TAGME* proposed by Ferragina and Scaiella [25]. TAGME is a tool to identify topics or short phrases in an unstructured or short text fragments. The topics in TAGME are identified from the hyperlinked texts in all the Wikipedia pages discarding ambiguous pages, list pages and redirect pages. It tries to find hypertext words in Wikipedia from the text and connect to them to a high related corresponding Wikipedia page. TAGME allows us to identify the context-based topics by understanding the text semantically. If we directly use terms from MedlinePlus dictionary, we can only capture the exact matching words in the sentence. Then if the sentence contains similar meaning but different words, we might miss the semantics in the text. However, if we use TAGME, we can obtain more semantically related topics in the text and find more accurate similarity between sentences. Table 2.1 shows an example of using TAGME. The sample sentence is the detailed information of term 'weight control' from MedlinPlus. As shown in the table, we find more related topics. Therefore, this tool is suitable to adopt for our topic identification method.

Table 2.1: Identifying topics using TAGME

Sentence	"Eating too much or not being physically active enough will make you overweight. To maintain your weight, the calories you eat must equal the energy you burn. To lose weight, you must use more calories than you eat."
Tagged Text	"***Eating*** too much or not being ***physically active*** enough will make you ***overweight***. To maintain your weight, the ***calories*** you eat must equal the energy you ***burn***. To ***lose weight***, you must use more calories than you eat."
Topics	**Eating Physical exercise Overweight Calorie Food energy Burn Weight loss**

Given the set of anchor texts $A(X)$ identified from a block of text X, the score for a particular sense p_x for the anchor text x to be associated with the page p is determined through a vote of all other anchor texts y which are in support of the annotation $x \xrightarrow{link} p$. Since the anchor text y can also have many senses, the vote is computed as the average relatedness for each sense p_y of the anchor y in relation to the sense p_x. Since not all senses of y have the same statistical significance, the contribution of p_y is weighted using its *commonness* or prior probability $Pr(p_y \mid y)$. Thus the voting formula is defined as:

$$vote_y(p_x) = \frac{\sum_{p_y \in G(y)} rel(p_y, p_x) Pr(p_y \mid y)}{\mid G(y) \mid} \tag{2.6}$$

where $rel(a, b)$ is a measure of relatedness between two pages a and b based on the overlap between their in-linking pages in Wikipedia. The relatedness score makes sure that only the senses p_y that are related to p_x affect the voting measure. The final score that defines the *goodness* of the annotation $x \xrightarrow{link} p$ is obtained by the sum of the votes of all other possible anchors y in the text T. The set of candidate anchors identified from the disambiguation phase are then passed through a pruning phase to discard possibly meaningless anchors. These bad anchors are identified based on the link probability of an anchor and the coherence of its candidate annotation which is computed as the average relatedness between the candidate sense of an anchor and the candidate senses for all other anchors in the given text. Only anchors with high link probability or whose assigned sense is coherent with the senses to other anchors are retained.

Chapter 3

Data Collection and Analysis

3.1 Data Collection

The first phase of the data analysis is to collect information about healthcare from the Internet for our analysis. First, we implemented a crawler in Python to download pages from the Web. The data was collected using a standard snowball approach, which follows the links in crawled pages to find new candidate pages. The crawl was done in parallel to maximize the use of resources. The initial seeds are sites of varying apparent quality chosen from the first few pages of arbitrary health-related Google searches. The seeds include government sites (e.g. nih.gov), university medical websites, hospitals, herbal remedy centers, etc. The seeds are chosen with an emphasis on diversity so that the crawl would quickly cover a wide range of sites, including both trustworthy and suspicious ones.

An early termination mechanism is enforced: a few initial pages crawled in a new domain are evaluated with a heuristic, which is based on a weighted set of approximately 150 health-related keywords. The heuristic is based on a weighted term frequency measurement, with words and phrases very roughly weighted based on their typical usage. For instance, “blood” is weighted low while “Blood pressure” carries a higher weight. We stop crawling a site if it fails the test. Although we do not further crawl the domain, we still keep existing and future links into the domain. Early termination is designed to keep the crawl on pages somewhat related to healthcare (regardless of quality). Domains failing this test are very unlikely to contain healthcare-related information, hence they are not crawled further. For instance, many websites contain a link to adobe.com, which directs users to download Adobe Reader or Flash Player. In this case, we do not want to further crawl the entire Adobe site. In practice, our approach selects a number of sites only tangentially related to healthcare, but also serves to both allow healthcare sites to be crawled, and to reject the majority of the irrelevant sites.

With the crawler, we have collected 316 thousand (316K) webpages from 39831 domains, with 3.4 million links between webpages. We further group webpages from the same domain (e.g. cdc.gov, nih.gov, who.int), and model the crawled network as a directed labeled graph.

In the graph, each node represents a domain (which contains all the pages from the domain), and each edge represents links between domains. An edge is labeled as the total number of links from the starting domain into the ending domain. In the generalized graph, the average number of links for a domain is 84.1. The maximum number of incoming links to a domain is 185,538, while the maximum number of outgoing links is 190,361.

For each link from site A to site B, if there also exists a link from B to A, we call them a pair of reciprocal links. The collected graph appears to be highly asymmetric – the reciprocal ratio is only 0.00955.

3.2 Link Distribution

First, we study the links between websites. In the Internet, a link from site A to site B is often regarded as an endorsement made by A in support of B, which is similar to the citation relationships in bibliography analysis. Figure 3.1 (a) shows the top 25 sites ranked by outgoing links and their corresponding incoming links; Figure 3.1 (b) shows the top 25 sites ranked by incoming links. We can see that healthcare-related information is not only

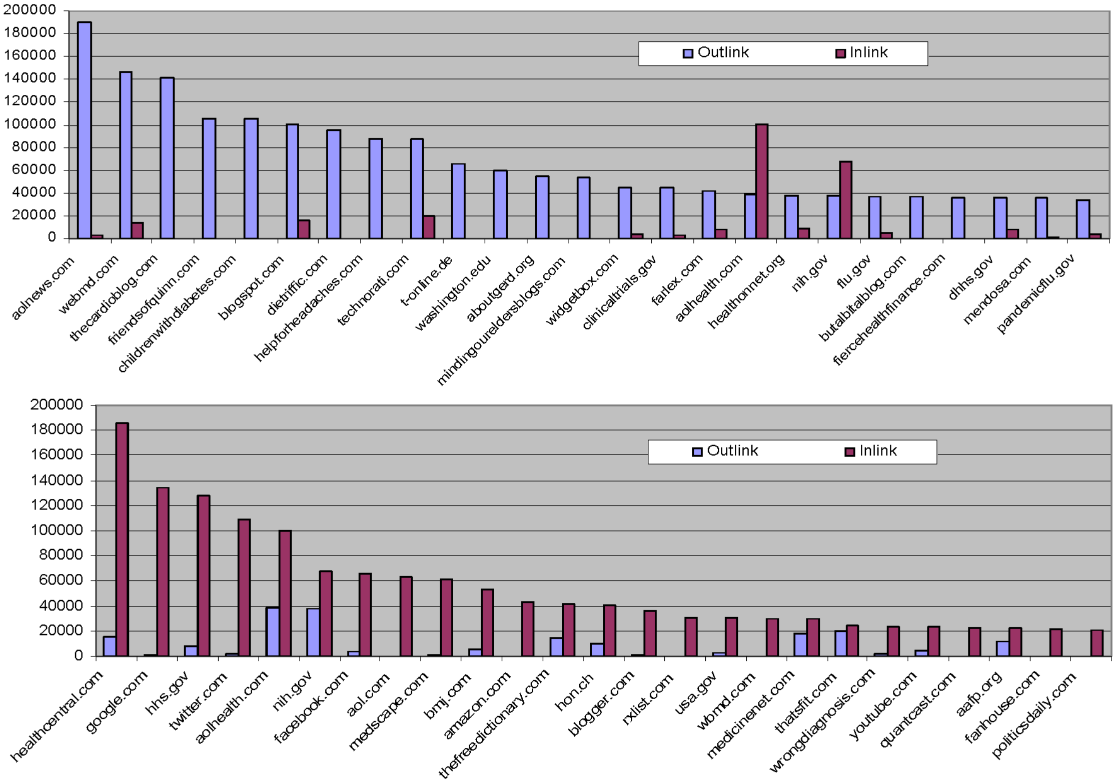


Figure 3.1: (a) Top 25 sites ranked by outgoing links; (b) Top 25 sites ranked by incoming links.

provided by professional websites. Rather, large amount of such information is published on general content providers (e.g. aol.com); or user experiences from social networking sites (e.g. twitter.com, blogger.com). We can also see that outgoing and incoming links are not symmetric: most of the sites ranked high in outgoing links does not have a large number of incoming links, i.e. the most active sites are not the most popular sites. As expected, health-related government agencies and large content providers rank higher in in 3.1 (b). This is consistent with users' perceptions of the most authoritative sites. Meanwhile, we have not observed strong link reciprocity (reciprocal ratio: 0.0096). In our data, site A linking to site B is very unlikely to result in site B linking back to site A. This is consistent with the measurements on the general web, but quite different from observations in online social networks.

We use snowball crawling to collect data - we follow links from crawled pages to access new pages. Because of this, it is natural that most of the sites post more outgoing links than incoming links. For a crawled site, we have identified all of its outgoing links, however, we only identified incoming links from other crawled sites, but not the entire web. We have also recorded a large number of terminal sites that we stopped following their outgoing links. This bias exists in all approaches using snowball crawling, since all known crawlers cover only a small portion of the Web (Note that this bias is not significant in social network analysis: due to strong link reciprocity, incoming links to a node mostly come from its neighborhood, and thus are easily collected through snowball crawling.). However, we can see from Figure 3.1 (b) that the top government agencies have many more incoming links than outgoing links, which means that their authoritativeness is widely acknowledged by other healthcare related sites (since we only crawled healthcare related sites), and their popularity surpasses their activeness.

Next, we study the overall distribution of links. Figure 3.2 shows the histogram of the number of outgoing links for each node. We can see a power-law distribution: a few nodes have a very large number of outgoing links (i.e. hubs); a moderate number of nodes have

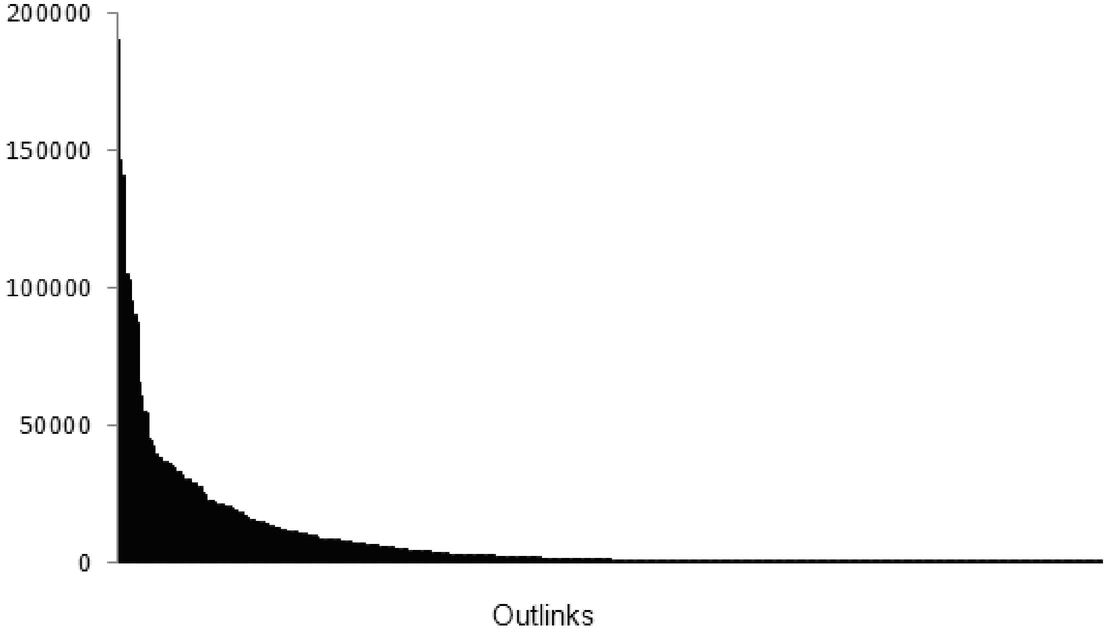
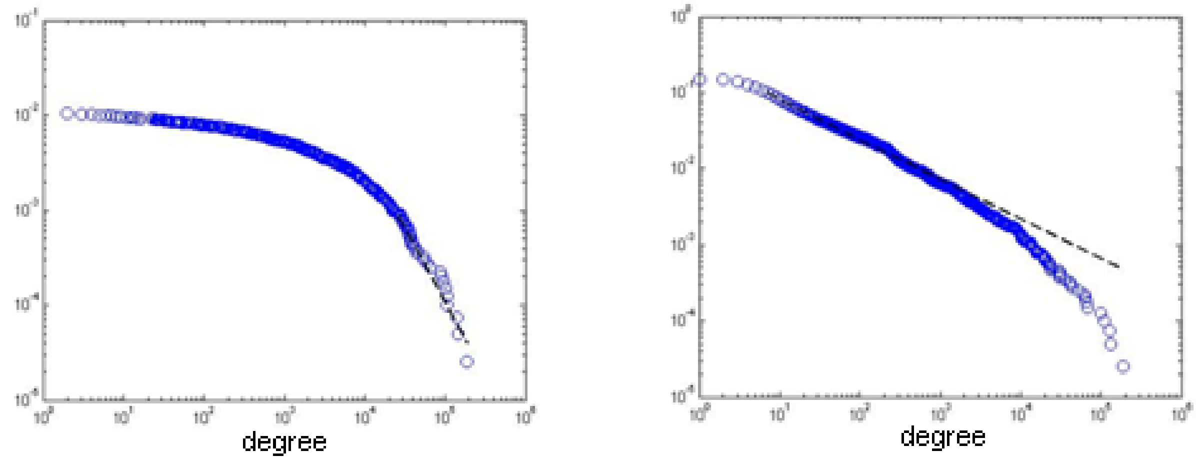


Figure 3.2: Distribution of out-going links.

	α	D
In-link	1.52	0.0263
Out-link	2.56	0.0986

Table 3.1: Power-law coefficient estimates (α) and K-S test metrics (D) for incoming and outgoing links.

a moderate number of outgoing links; and a very large number of nodes have very few outgoing links. This appears to be consistent with the measurements over the general Web, as well as various social networking graphs. To confirm this observation, we further test the graph structure of using the method proposed in [9]. The method uses maximum-likelihood estimation and Kolmogorov-Smirnov goodness-of-fit metric to calculate the best power-law fit. We plot the complementary cumulative distribution functions (CCDF) in Figure 3.3, and find that the distribution of outgoing links and incoming links both follow the power-law, which also satisfy scale-free network property. The power-law coefficient estimates (α) and Kolmogorov-Smirnov test metrics (D) for both distributions are shown in Table 3.1.



(a) (b)

Figure 3.3: Log-log plot of complementary cumulative distribution functions for: (a) incoming links; (b)outgoing links.

3.3 Domain Relationships

As we have addressed in 1, for online healthcare information, the credibility of information and the trustworthiness of the information sources are very important but relatively difficult to measure. A very coarse but generally accepted understanding is that .gov sites carry relatively scientific and reliable information, while the credibility of .com sites are somewhat mixed. Therefore, to further understand the link distribution of healthcare information providers on the Web, we employ a summarization approach to categorize our graph based on top domains: (1) .gov is restricted to government entities, and this restriction is enforced. (2) .edu is designed for post-secondary institutions and organizations, and this restriction has been enforced since 2001. (3). .com is designed for commercial use, and is publicly available. On the other hand, although .net was intended for network-related organizations (e.g. ISPs), this has never been enforced, and .net is now treated as an alternate to .com. Hence, we merge .com and .net sites. (4) .org is designed for non-profit or non-commercial organizations, but this restriction has also never been enforced. In practice, it is primarily used by the intended consumers, but it still used by a diverse group of organizations. (5) we

Domain	Sites	Pages	Internal Links
.com	25852	177437	28202
.org	7904	43921	3778
.gov	516	53429	1566
.edu	613	23450	330
others	4946	17890	1016

Table 3.2: Node and link distribution among top-level domains.

group all other top domains into the last category.

Table 3.2 shows the distribution of domains, pages, and internal links (i.e. links starting and ending in the top-level domain) in each category. Note that we did not include links starting and ending in the same sites (e.g. from a page in hhs.gov to another page in hhs.gov). Table3.3 further presents the statistics for each category. The link density is a global measure measurement, which is defined as the proportion of existing links over total number of links possible in the (sub)graph. To further measure the tightness of connections in a local neighborhood, we measure the average clustering coefficient in each category. The clustering coefficient of a node with N neighbors is defined as "the number of directed links that exit between the node's N neighbors, divided by the number of possible directed links that could exist between the node's neighbors (i.e. $N \times (N-1)$)". We calculate the clustering coefficient of a domain by the average clustering coefficient of all nodes in the domain.

From the table, we can see that, although .com contains the greatest number of pages and links, the nodes in this domain are not as inter-connected as other top-level domains. The link density and clustering coefficient are both low. However, density and clustering coefficient from the .gov and .edu domain are the highest. This is partly due to the early termination mechanism in crawling: many of the .com and .org sites are terminals (i.e. non-health-related sites that we do not further crawl), hence they do not further contribute to links.

Furthermore, we study the distribution of outgoing and incoming links by domain. Figure 3.4 demonstrates the distribution of outgoing links from each top-level domain to all categories. Please note that the number of links in each pie piece is normalized by the num-

Domain	Link Density	Clustering Coefficient
.com	0.08E-03	0.39E-03
.org	0.12E-03	1.01E-03
.gov	11.7E-03	8.44E-03
.edu	1.75E-03	1.17E-03
others	0.08E-03	0.44E-03

Table 3.3: Statistics for top-level domains.

ber of pages in the destination domain. Therefore, the pie-chart represents the distribution of outgoing links with respect to the size of the destination domain. Figure 3.5 shows the distribution of incoming links to each top domain from all categories, normalized by the size of the origination domain. From the figure, we can see that .org is the most popular destination, while .com appears to be unfavorable except by itself. In fact, all to-level domains post significant numbers of links to .com, however, since .com is the largest category, the proportions become small. Unexpectedly, .gov is not a popular destination either. With further analysis of the data, we found that government sites are relatively large (on average), but many of the external sites only have a link to the front page.

To highlight the primary contributors from/to each domain, we exclude all minor pieces (< 10%) from the pie chart, and redraw the remaining in Figure 3.6 (note that we eliminated "others" category). From the figure, we can see that the link between .gov and .com are quite weak – the are on the two ends of the spectrum. Meanwhile, due to the fact that .org domain is relatively diversified (contains both highly trustable organizations, as well as highly suspicious personal or small-size commercial users), it is somewhat balanced in terms of link distribution. .edu is similar to .org. We expected .edu to show stronger association to .gov, however, the observation disproved our prediction.

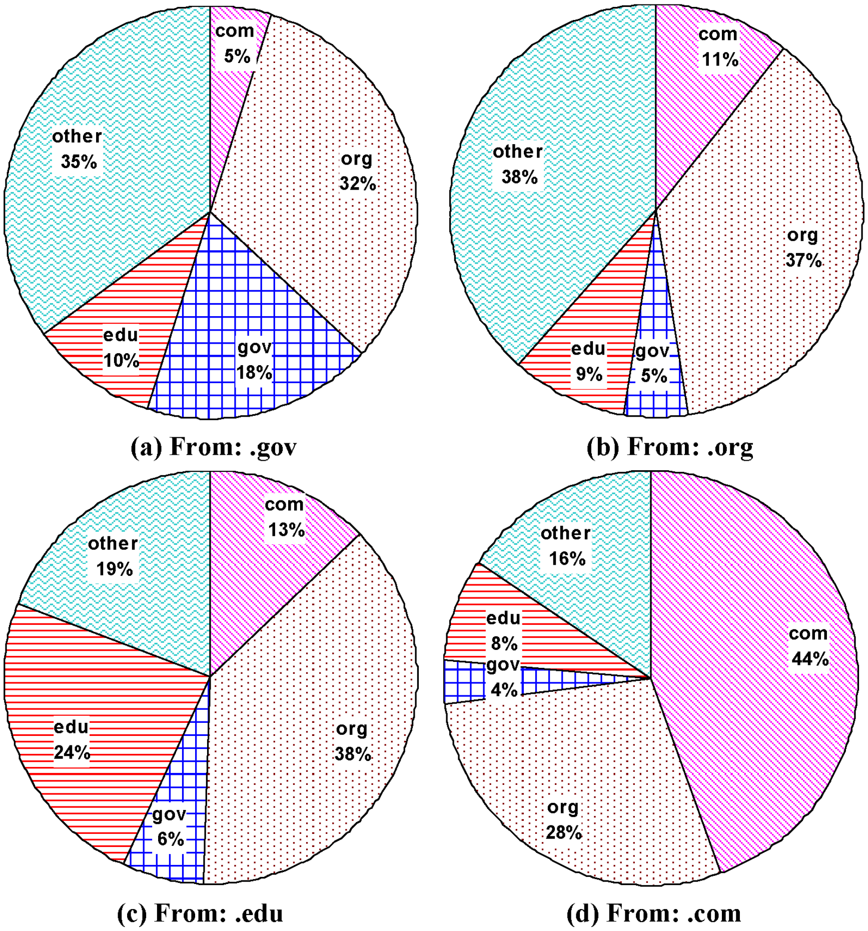
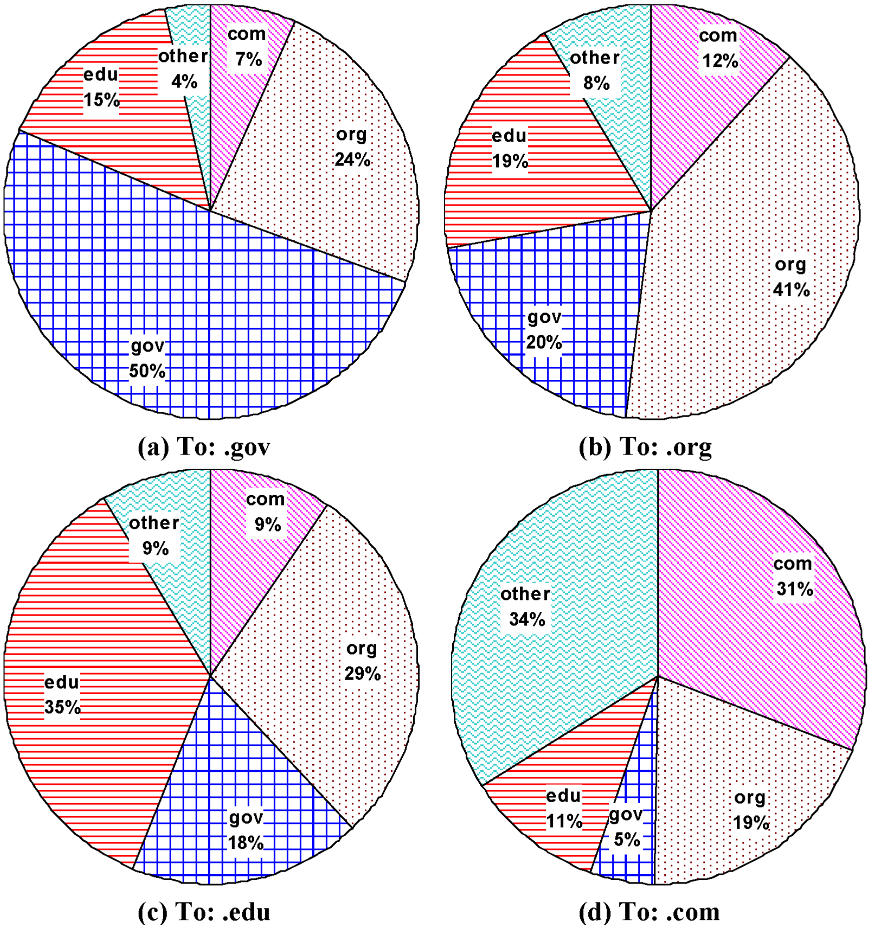


Figure 3.4: Distribution of outgoing links by domain.



(a) To: .gov

(b) To: .org

(c) To: .edu

(d) To: .com

Figure 3.5: Distribution of incoming links by domain.

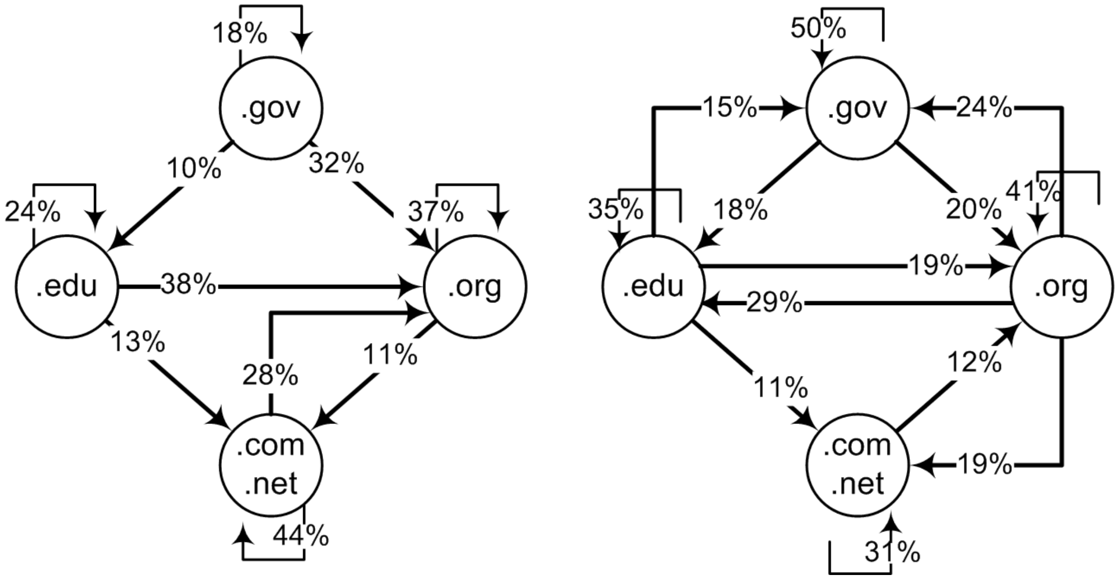


Figure 3.6: Outgoing and incoming links.

3.4 User study

3.4.1 Experiment I: Search

In the first user study, we designed 10 queries related to healthcare, and asked experts to judge the trustworthiness of the top 20 results returned from a commercial search engine: Google. The queries are:

- "stomach flu treatment"
- "chiropractic massage therapy"
- "emergency treatment"
- "EKG" (electrocardogram)
- "cushing's syndrome"
- "menopause"
- "exercise muscle"
- "diet weight loss"
- "sprain treatment"
- "norovirus"

The goal is to study the reliability of the search engine, especially if it returns credible results. Overall, 51% of the sites are labeled as "credible" by the users, while 45% of the sites are found to be "suspicious", and the other sites (4%) are irrelevant or inaccessible. A breakdown of the labels is shown in Figure 3.7.

As we can see from the figure, for every query, there are at least a few suspicious sites returned by the search engine. Again, it proves that search engines do not use trustworthiness in their scoring mechanism. The mixed results could be very confusing to the users. Although

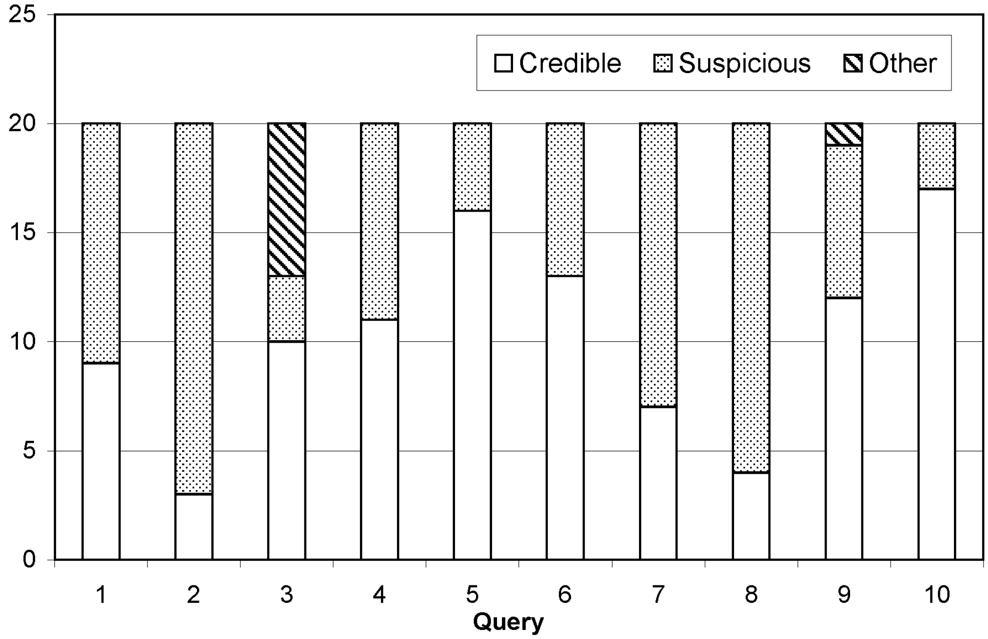


Figure 3.7: Credibility of healthcare-related search results judged by users.

our experts are capable of distinguishing trustworthy and untrustworthy sites, it could be a difficult task for regular users, especially consider that the Internet is currently used by a very diverse population. To confirm this, we have designed another user study.

3.4.2 Experiment 2: Search Result

In this experiment, we search for "heart attack medicine" using Google, one of the top 10 results is a particularly suspicious webpage. It contains 45 sponsored advertisements, most of which sells herbal or dietary supplements that are not FDA approved. FDA evaluation and approval is not required for such products, however, only a few of these websites properly contain the FDA-required disclaimer that such products "are not intended to treat, diagnose, or cure any disease." In a user-based evaluation, we have asked 22 participants (undergraduate and graduate students, faculty members) to judge the trustworthiness of the webpage. Among the 22 responses, only one of them thinks that the webpage is trustworthy and the content is scientific and authoritative. Meanwhile, a large portion of the users (68%) think that the webpage contains both trustworthy and false contents, or cannot determine the trustworthiness of the webpage. On average, it takes 188 seconds for a user to make the verdict. It appears that the webpage – although ranked highly by Google – confused many of the participants. They took a long time to determine the trustworthiness of the webpage, and many of them still could not make a judgment.

Web users are highly diverse - not all of them have the expertise to judge the correctness and trustworthiness of all of such websites. In our experiment, all participants are highly educated with various backgrounds (CS, EE, Biology and Chemistry). However, many of them still had difficulty with the task. Meanwhile, when participants are introduced to a potentially dangerous combination of information that appears to be correct together with information that is extremely suspicious, most of them found it difficult to judge the credibility of the information source.

The results of the user studies demonstrates the very mixed quality and credibility of online healthcare information. Many people now use the Internet as their primary information source. However, our results show that users are very vulnerable when they seek for health advice and information over the Internet.

Chapter 4

Seed Set Selection

Seed selection is very critical in ranking algorithm to assess the trustworthiness of the webpages. TrustRank proposed a Inverse PageRank to identify the initial seed set. However, it needs human effort to curate whether the seed set is trustworthy or not. In this proposal, we use automatic seed set selection method using Affinity Propagation clustering method to exclude manual work. Section 4.1 show the Inverse PageRank process for TrustRank and followed section describes the seed selection method for our proposed system, HealthTrust.

4.1 Selecting seeds from Inverse PageRank

As described inInver PageRank in Chapter 2 is taking opposite direction of the link structure of the Web graph. Since the behind logic of the PageRank is that the important sites have more incoming links than relatively unimportant sites, Inverse PageRank finds the important site from the inverse structure which has more outgoing links. We construct the inverse web graph and perform the PageRank with the link structure. Table 4.1 is the Top 15 seeds from the Inverse PageRank. As seen in the table, 'healthcentral.com' and 'hon.ch' are the most popular websites among healthcare related websites on the Internet. However, we know that popularity does not always mean the trustworthiness as we mentioned in Chapter 1. Furthermore, several websites such as 'blogspot.com', technorati.com' and 'apple.com' are not trustworthy. Therefore the manual curating process is inevitable for selecting seed set in TrustRank process.

Table 4.1: TrustRank Seed Set

Domain	Rank
healthcentral.com	0.0645171
hon.ch	0.0501275
blogspot.com	0.0297561
childrenwithdiabetes.com	0.0174915
mendosa.com	0.0170831
healthonnet.org	0.0166101
netwellness.org	0.0164902
cdc.gov	0.0163779
diabetesmonitor.com	0.0144608
nih.gov	0.0143085
technorati.com	0.0114687
healthscout.com	0.00989544
ic-network.com	0.00947256
apple.com	0.00932392
familydoctor.org	0.00897219

4.2 Selecting seeds from AP Clustering

As described in 2, AP clustering finds exemplars among nodes and propagates the affinity to form optimal clusters in the network using similarity measure. Since the similarity measure is not necessarily to be an Euclidean distance and symmetric, we have tested two similarity measure in HealthTrust. Also we have changed the preference values for each case to get optimal clusters. The preference is a priori knowledge of how good the node is as a center, therefore, in our method, we have used two different cases as preferences. However, if the value of preference is zero, we have replaced it with 0.1 and 0.01 respectively, considering that the node might have a link from or to some other nodes from web network. We have defined the similarities and preferences as below.

Similarity:$S(i, j)$

- Asymmetric similarity: the number of links from node i to node j.
- Symmetric similarity: sum of the number of links from both directions, *i.e.* $S(i,j) = S(j,i)$.

Preference:$P(i)$

- The number of domains that point to node i.
- The average number of links between other nodes.

Since the AP algorithm tries to find the maximum net similarity, we have chosen the *Asym-p01-no-terminal-node* to consider for our seed sets. However, the symmetric cases get more clusters and the number of elements are much bigger than the asymmetric cases. Thus, we have discarded the results from symmetric cases. In addition, when we have used the number of domains as a preference value, we have found more optimal clusters. We would like to see the effect of the dangling nodes, we have evaluated the case that excludes them. We found the optimal clustering result with maximum similarity in the case; Asymmetric similarity measure and the number of domain preference without terminal nodes. The results are shown below in Table 4.2 excluding the other cases. Table 4.3 shows the examples of cluster we have found and the elements of NIH cluster is given in Table 4.4.

Ranking algorithm using seed sets is very sensitive to selecting proper seed sets. We tested modified TrustRank with manually selected trustworthy sites such as “nih.gov” to see how many trustworthy sites can be ranked within top 20. The results are shown in Figure 4.1. It is clear that the more reliable sites are included, the more information sites are from authoritative sites. However, AP clustering also gives us many clusters and hubs containing small number of elements. Due to the nature of Web graph, many websites just form its own cluster with an element which makes a lot of isolated clusters from the clustering. Even among the authoritative sites, they are divided into different clusters. To address this issue, we need a better selection method for seed sets.

Table 4.2: AP clustering Results

Similarity Type	Preference	Net similarity	Number of Clusters (size>=2)	Number of Clusters (size>=3)	Number of Clusters (size>=5)
Asym-p0	domain	$1.6537e+006$	176	24	11
Asym-p01	domain	$1.6569e+006$	177	25	12
Asym-p001	domain	$1.654e+006$	176	25	12
Asym-p0	link	$1.6419e+006$	223	28	12
Asym-p01	link	$1.6419e+006$	222	28	12
Asym-p0-no-terminal-node	domain	$1.6537e+006$	171	24	10
Asym-p01-no-terminal-node	domain	$1.6569e+006$	172	24	10
Asym-p001-no-terminal-node	domain	$1.654e+006$	170	25	10
Asym-p0-no-terminal-node	link	$1.6419e+006$	204	26	10
Asym-p001-no-terminal-node	link	$1.6419e+006$	205	26	10
Sym	domain	$3.2879e+006$	233	201	182
Sym-no-terminal-node	domain	$3.2879e+006$	215	192	168

Table 4.3: Example of clusters

Center site	Number of elements in the cluster
adobe.com	7
hon.ch	14
nih.gov	9
twitter.com	12
usa.gov	5
healthcentral.com	15
facebook.com	6
whitehouse.gov	7
feedburner.com	5
google.com	15

Table 4.4: 'hhs.gov' cluster

Hub domain	Elements in the cluster
hhs.gov (U.S. Department Health & Service)	ahrq.gov cdc.gov fda.gov flu.gov insurekidsnow.gov cms.gov foodsafety.gov health.gov hrsa.gov medicare.gov womenshealth.gov pandemicflu.gov samhsa.gov childwelfare.gov phe.gov

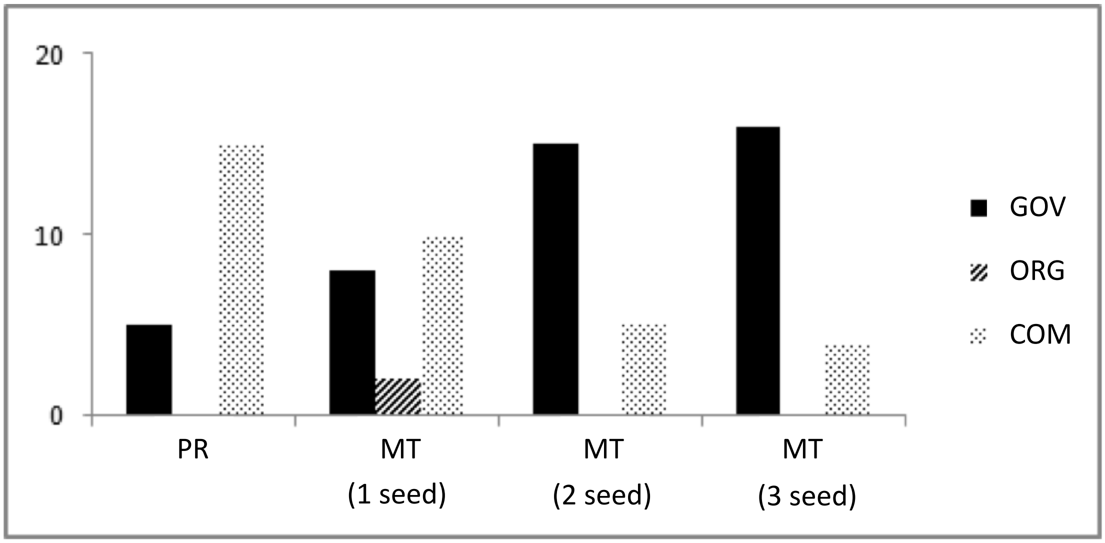


Figure 4.1: Top 20 with different number of seeds

Chapter 5

Analysis for the Healthcare Information Web

We have experimented our data set for three ranking algorithms, PageRank, HITS, and TrustRank respectively. The results show that ranking algorithms are not enough to identify the trustworthy websites as a sole method. Additionally, we evaluated the term frequency of the randomly chosen websites from credible and suspicious websites. The results are shown in section 5.2.

5.1 Ranking Algorithms

5.1.1 PageRank

We applied PageRank on the healthcare information network that we collected. Results are shown in Table 5.1. As we can see, the results are very mixed. Many non-health related websites are ranked very high, since they are frequently pointed-to by healthcare related websites. Meanwhile, social networking sites, such as Twitter, Facebook and Blogspot are ranked very high – this may introduce a potential risk that personal experiences or advices from social networking sites be ranked high in a healthcare-related search. When we further

Rank	**Name**	**PageRank Score**
1	healthcentral.com	0.0032786
2	twitter.com	0.0030572
3	facebook.com	0.0026880
4	google.com	0.0023624
5	thefreedictionary.com	0.0021416
6	adobe.com	0.0020395
7	*hon.ch	0.0014591
8	youtube.com	0.0014153
9	*cdc.gov	0.0012058
10	*nih.gov	0.0012026
11	farlex.com	0.0011012
12	shopwishlist.com	0.0010909
13	blogspot.com	0.0009334
14	thehealthcentralnetwork.com	0.0008741
15	wikipedia.org	0.0008634
16	*hhs.gov	0.0008625
17	usa.gov	0.0008506
18	cafepress.com	0.0007953
19	thefreelibrary.com	0.0007692
20	definition-of.com	0.0007687

Table 5.1: Top 15 PageRank results

look into our data, we found that we do have a large number of healthcare-related pages from blog space, which give personal advices, and refer to a mixture of reliable and suspicious sites. Meanwhile, some highly credible government sites and non-profit noncommercial organization sites are also ranked very high.

5.1.2 HITS

Another popular link analysis approach is HITS [**?**] described in Chapter 2, which divides webpages into “hubs” (pages with many outgoing links) and “authorities” (pages with many incoming links). It is based upon the idea that “good hubs” point to “good authorities”,

Rank	Authority Name
1	twitter.com
2	google.com
3	youtube.com
4	facebook.com
5	nih.gov
6	wikipedia.org
7	nytimes.com
8	yahoo.com
9	adobe.com
10	apple.com
11	blogspot.com
12	cnn.com
13	amazon.com
14	cdc.gov
15	fda.gov
16	flickr.com
17	washingtonpost.com
18	wordpress.com
19	about.com
20	go.com

Rank	Hub Name
1	blogspot.com
2	cdc.gov
3	wikipedia.org
4	nih.gov
5	usda.gov
6	washingtonpost.com
7	typepad.com
8	nytimes.com
9	clinicaltrials.gov
10	usatoday.com
11	cnn.com
12	google.com
13	twitter.com
14	go.com
15	businessweek.com
16	webmd.com
17	yahoo.com
18	harvard.edu
19	ama-assn.org
20	youtube.com

Table 5.2: Sites with top authority and top hub scores.

while "good authorities" are pointed to by "good hubs". Hub and authority scores are thus calculated through an iterative approach.

We implemented the HITS algorithm, and applied it on the collected of healthcare information network. Results are shown in Table 5.2. Again, the result is very mixed. In particular, the top 4 authorities are all social networking and search engines – many sites have a link to these sites, pointing to a YouTube video, a Facebook group, or a user on twitter. On the other hand, the most reliable healthcare sites such as nih.gov and cdc.gov are also ranked highly, which correctly represents their authoritative status.

Table 5.3: Top 15 TrustRank results

Name	**Score**
healthcentral.com	0.141919
hon.ch	0.140551
twitter.com	0.006129
facebook.com	0.005049
adobe.com	0.004185
google.com	0.004152
nih.gov	0.004081
youtube.com	0.003625
thehealthcentralnetwork.com	0.00308
blogspot.com	0.002639
hhs.gov	0.002561
addtoany.com	0.002192
clinicaltrials.gov	0.002132
yahoo.com	0.002087
apple.com	0.001905
ftc.gov	0.001891
healthscout.com	0.001886
go.com	0.001873
foodfit.com	0.001872
cdc.gov	0.001871

5.1.3 TrustRank

Next, we performed TrustRank with 2 seeds which is obtained from top 15 websites after running Inverse PageRank. As shown in Table 4.1, the seed set contains trustworthy sites and commercial sites as well. The seed set should manually curated before used in TrustRank, therefore we only included two seeds after manually curated – *healthcentral.com, hon.ch* – to evaluate the accuracy of the TrustRank. Basically, TrustRank propagates the credible sites based on the seed set. However, the TrustRank results in far less number of authoritative sites within top 20 and unrelated sites such as 'adobe.com' and 'apple.com' are ranked high within top 20.

5.1.4 Modified TrustRank

After identifying the seed set automatically using AP clustering as shown in 4.2, we performed our modified TrustRank algorithm. Modified TrustRank is improved from TrustRank algorithm considering the number of outgoing links for each site. TrustRank only modify the teleporting terms of the PageRank to propagate the trustworthy pages. However, since we construct the domain-based web graph, we take into account the number of outgoing links from node A to node B. Considering the number of outgoing links, we can weight the importance of a node relatively. Modified TrustRank can be used for ranking each page as well. We perform our ranking algorithm to get the trustworthy scores for assessing the credibility of the domains. Since we define $\mathcal{F}(u, v)$ which is the number of links from node u to node v. So our modified TrustRank is computed by Equation 5.1.

$$\mathcal{DR}(u) = \alpha \cdot \sum_{v:(v,u)\in E} \frac{\mathcal{F}(v, u)}{\omega(v)} \cdot \mathcal{DR}(v) + (1 - \alpha)\mathbf{d} \tag{5.1}$$

Modified TrustRank gives high ranks for the authoritative domains such as government or organization sites so that the one can rely on their searched results in which biased or commercially recommended information is excluded. The results of modified TrustRank have been compared with the original PageRank and TrustRank. We used seeds from "hhs.gov" cluster, which is 15 seeds, for modified TrustRank. The information sources are ranked based on the trustworthiness scores returned. Top 10 sites excluding seeds are listed in Figure 5.4. As seen from the figure, modified TrustRank gives all authoritative sites within top 10. Furthermore, in order to compare the results with PageRank and TrustRank, we fix *nih.gov, hhs.gov* as two seeds. In Table 5.5, the results show that definitely modified TrustRank outperformed compared to other ranking algorithm, PageRank and TrustRank. However, if we reduce the number of seed set, suspicious sites such as 'thebody.com' are ranked very high. In oder to overcome this issue, we propose a content-based method for HealthTrust in order to consider the semantics of the resources and it is explained in section 6.3.

Table 5.4: Top 10 DomainTrust results with *hhs.gov* cluster

Name	**Score**
samhsa.gov	0.062481
childwelfare.gov	0.062481
smokefree.gov	0.000057
epa.gov	0.000045
amia.org	0.000045
letsmove.gov	0.000032
aafp.org	0.000031
himss.org	0.000031
medlineplus.gov	0.000012
aafa.org	0.000011

Table 5.5: Comparison with PageRank, TrustRank and DomainTrust

Rank	**PageRank**	**TrustRank**	**DomainTrust**
1	twitter.com	nih.gov	nih.gov
2	facebook.com	hhs.gov	hhs.gov
3	healthcentral.com	usa.gov	epa.gov
4	yogawiz.com	adobe.com	smokefree.gov
5	google.com	cdc.gov	amia.org
6	hon.ch	whitehouse.gov	letsmove.gov
7	adobe.com	twitter.com	unc.edu
8	youtube.com	youtube.com	washingtonpost.com
9	usa.gov	fda.gov	about.com
10	nih.gov	facebook.com	aboutgerd.org
11	thehealthcentralnetwork.com	medlineplus.gov	thebody.com
12	blogspot.com	usda.gov	alzfdn.org
13	cdc.gov	microsoft.com	expasy.org
14	hhs.gov	flu.gov	reflux.org
15	digg.com	gpo.gov	flu.gov
16	feedburner.com	medicare.gov	aanma.org
17	doubleclick.com	usdoj.gov	asph.org
18	ftc.com	dhhs.gov	aolnews.com
19	delicious.com	opm.gov	clinicaltrials.gov
20	addtoany.com	ahrq.gov	nof.org

5.2 Term Frequency

Term Frequency (TF) is commonly used weighting scheme in information retrieval to characterize a document by counting the number of occurrences of the term in the document. Mostly, TF is combined with Inverse Document Frequency (IDF) for relatively weighting the term occurrences in a corpus, which is called TF-IDF. TF-IDF is calculated using the Equation 5.2 : the number of times the term appeared in a document (Term Frequency) multiplied by the rareness of the term across all documents in a corpus (IDF). TF-IDF provides the distribution of terms in a corpus weighting the frequency of the terms. Therefore unimportant terms or strop word such as ''the' can be filtered out. Initially, simple ranking systems only calculate the TF-IDF for the query since the more the query appears in a document, the more the document is related.

$$\begin{aligned} tfidf(t,d,D) &= tf(t,d) \times idf(t,d,D) \\ tf(t,d) &= \frac{f(t,d)}{max\{f(w,d):w\in d\}}, \\ idf(t,d,D) &= \log\frac{|D|}{|\{d\in D:t\in d\}|} \end{aligned} \tag{5.2}$$

where t is the term, d is a document, and $|D|$ is the total number of documents.

In this proposal, we evaluated the distribution of term frequency between credible and suspicious sites. First, we randomly selected six credible sites from authoritative sites and six suspicious sites from commercial sites. We calculated the difference of credible and suspicious TF values, $tf(Good) - tf(Bad)$. Figure 5.1 and 5.2 show the distribution of top 20 TF terms for $tf(Good) - tf(Bad)$ and $tf(Bad) - tf(Good)$. As seen in the figures, the top 20 terms are very different from each case. Credible sites use more general terms related to health and suspicious sites tend to use more specific terms related to their websites' products. However, we cannot say the sites are untrustworthy because it is possible that they use their own terms in their site for advertise their products. In addition, only using TF or TF-IDF has a limitation for ranking problem, since it doesn't consider any semantics at all and only dependents on the frequency of the terms occurred in a document. Therefore we need a

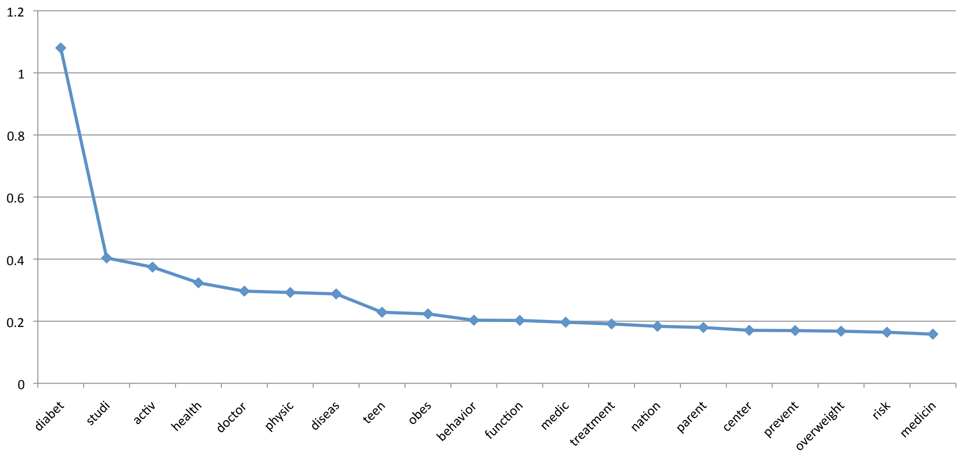


Figure 5.1: Top 20 terms from $tf(Good) - tf(Bad)$

further study to analyze the semantics for trustworthiness.

In this preliminary work, a thorough study over a large volume of online healthcare information websites collected by a focused crawler is presented. With a focus on the providers of such information, questions such as who they are, how they are distributed, and how they are related were answered. The network structural features, analyze the graph topology, and study the nodes and links distributed over top level domains are also measured. Two link analysis approaches, PageRank and HITS to study the authoritativeness of websites based on graph topology are used. As the results are shown, traditional approaches give mixed results of credible, suspicious, and irrelevant sites.

With two user studies, commercial search engine results for health-related queries are far from satisfactory and many untrustworthy or highly suspicious sites are returned. Meanwhile it is not easy for users to distinguish trustworthy sites from the mixed results. A reliable

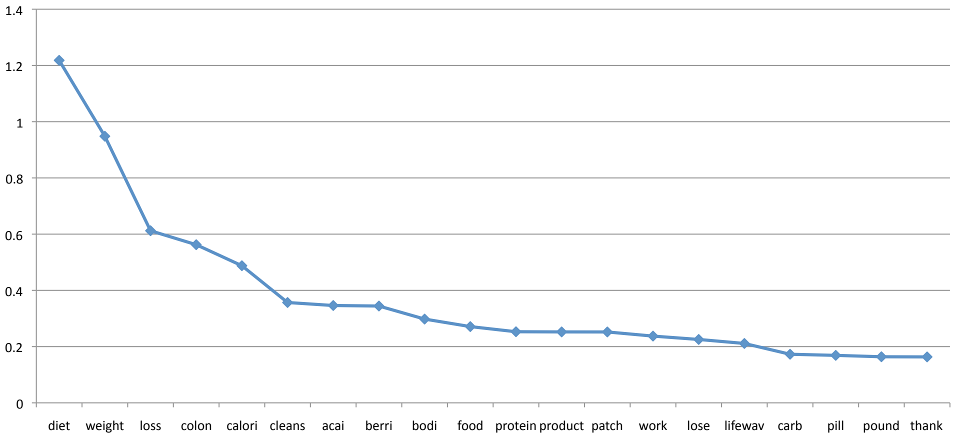


Figure 5.2: Top 20 terms from $tf(Bad) - tf(Good)$

mechanism to automatically determine the trustworthiness of online healthcare information is highly desired.

The next step of the project is to implement such a mechanism. Ideally, it will assess the credibility of online healthcare information sources by evaluating their topological relationships and content similarities with trusted websites. However, it is still highly challenging to assess content similarity at semantic level. Meanwhile, another interesting future direction is to further understand healthcare information consuming behaviors of the users.

Chapter 6

The HealthTrust System

The fundamental goal of our research is to assess the trustworthiness of healthcare information on the Web. As well recognized, healthcare information is growing exponentially along with the web technologies such as social networking services, real-time web technologies such as wikis, blogs, and RSS (Really Simple Syndication). However, not all health information provided online is trustworthy. Although trustworthiness of the online healthcare information is essential for a great number of daily users, not much work has been done in determining the trustworthiness of these sources so far. In this section, we propose a new methodology, called HealthTrust. The HealthTrust aims to identify the trustworthy information automatically in the flood of healthcare related data on the Web and provide more credible guidance to consumers.

The overview of HealthTrust framework is introduced in 6.1. Section 7.1.1 describes the data set we will use for the proposed research. More explanation of each component of the HealthTrust is given in Section 6.2 and 6.3 in detail.

6.1 Overview of the HealthTrust

The HealthTrust algorithms consist of three phases; *Structure-based Analysis*, *Content-based Analysis* and *Integration.* The two approaches are integrated for ranking a final HealthTrust

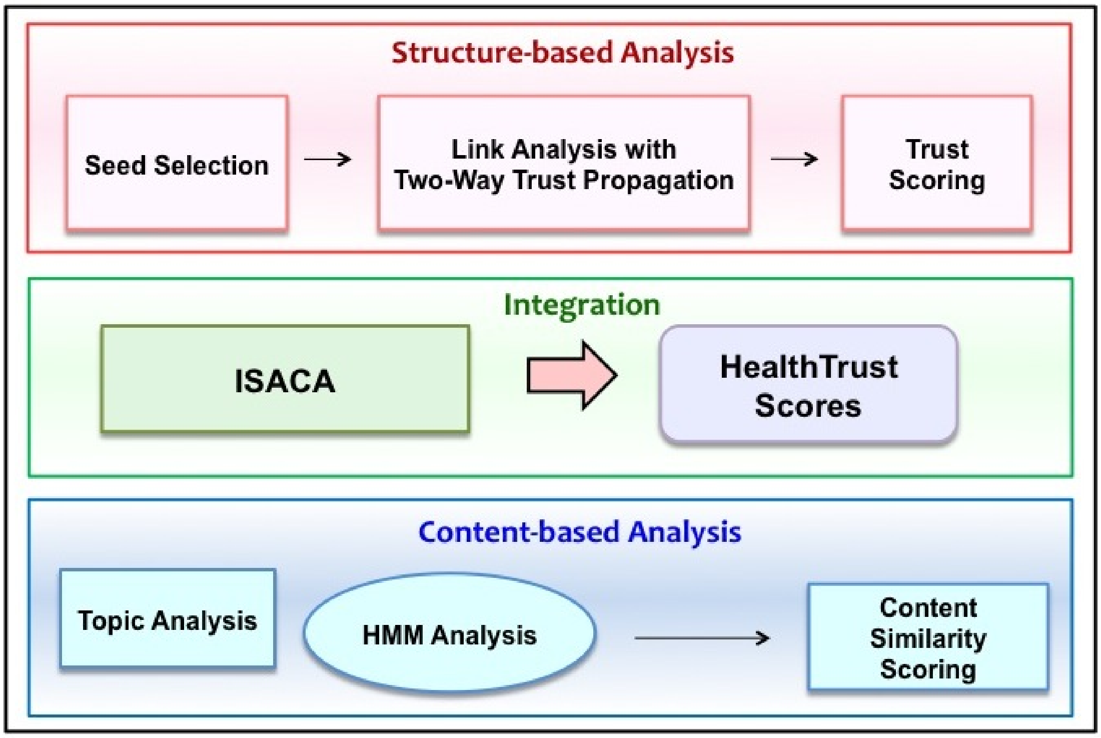


Figure 6.1: Overview of the HealthTrust

score for each node. Figure 6.1 shows the overall system structure of HealthTrust. In the first phase, Affinity Propagation (AP) [27] clustering is adopted to find the seed set automatically. Such clusters will be easily identifiable and used as (positive and negative) seeds in the future steps. Based on the seed set, our new ranking algorithm is performed to propagate the trustworthy information and produces the ranking scores for each node. Next, in the second phase, the consistency of contents between healthcare-related webpages and trusted sites is evaluated. Due to the lack of semantic analysis for traditional topic modeling based on the frequency of the words in a document, we conduct two novel approaches, Topic Analysis and Hidden Markov Model analysis. Using two methods, we would like to identify the trustworthiness of the contents. As a result of the both analysis, content similarity scores are generated for each method. Then, we get a final verdict - a HealthTrust score for each node integrating structure-based and content-based analysis with an iterative way. In the following sections, we describe each component of HeathTrust framework in more detail.

The related research for ranking websites has commonly used the structural information such as url, domain names, xml properties, and link structure. However, existing methods are not suitable to extract the meaningful information from unstructured data such as free-style texts, tweets, or comments of Facebook. Since we take into account the semantic analysis using topic modeling techniques and integrate with the structural ranking analysis, our algorithms are more reliable and reasonable than the existing methods.

6.2 Structure-based Analysis

We use the seed sets obtained from AP clustering. As we mentioned earlier in the Chapter 4, selecting seeds is sensitive and affects the ranking algorithm. However, AP clustering has fundamental drawbacks. First, AP clustering may produce too many clusters containing small number of elements. Since some nodes have only small number of links they refer, the nodes can be isolated from other nodes. As a result, they form their own cluster with one element which result in a lot of isolated clusters. Secondly, the authoritative sites may be divided into different clusters. For example, if 'nih.gov' and 'cdc.gov' are found in different clusters, we may miss an important hub by choosing one of the clusters. To address this issue, we need a better way to select seed sets.

In this research, we simply select hubs from the clusters. Based on the AP clustering, hubs or exemplars have the maximum affinity between other elements in the cluster. It means that hubs play an important role in clusters, thus, hubs themselves can be a seed. We make two groups of seed sets as *positive* seed set and *negative* seed set and assign the same number of hubs into the two groups. However, if the cluster only contains one or two nodes, we discard them since the nodes are not effective to form a cluster. Therefore we only consider clusters with more than three elements.

After we get the seed sets, we perform a new ranking algorithm to propagate the trustworthy websites. As seen in the Chapter 5.1.3, TrustRank performs better than the PageRank, but still it contains many suspicious sites due to its seed set. Our new automatic seed set selection method overcomes the problem. Another problem is that TrustRank is too sensitive to propagate the trustworthy websites than we expected. If the graph becomes greatly enlarged, then the speed of propagation might not be satisfactory. Hence we use a new ranking algorithm improved from TrustRank.

With a new ranking algorithm, called TwoWay-TrustRank, we can start a parallel propagation. Since we have two seed sets, positive and negative sets, TwoWay-Rank propagates to the credible and suspicious nodes in a graph simultaneously. we expect the algorithm to be less sensitive and produce accurate ranking scores with faster execution time. In order to implement this method, we define the teleporting factor **d** in a different way. We don't start $d = 1$ for seed set. For the positive seed set, the teleporting factor is a positive number, and for negative seed set d is a negative number, for instance, 1 and -1 respectively.

6.3 Content-based Analysis

We want to identify the veracity or trustworthiness of the information provided in healthcare sites by performing content analysis using a Hidden Markov Model (HMM) and Topic modeling. Our approach is to be able to classify the information available in the healthcare domain into two categories:

- trustworthy websites like www.health.gov and other similar .gov or .edu sites
- suspicious sites which often end with a .com as a part of their domain name

In order to differentiate the content available in these two categories of websites, we need to examine the information presented in these sites. In general, the information presented in websites tend to be a combination of both trustworthy and suspicious information. Since a sentence is a more fundamental unit for presenting information, we plan to perform the analysis at the level of the sentences available in these websites. In our approach, we have built two HMMs: one to model a trustworthy sentence from a trustworthy site and another to model a suspicious sentence from a suspicious website. The models trained on sample sentences from the good and bad websites respectively can then be used to evaluate candidate sentences to identify if they are similar to the trustworthy or suspicious sentences. All candidate sentences in a given web page can be evaluated similarly and then an aggregate measure or a classifier can be used to determine if the content of a page is trustworthy or not. Similarly an aggregate measure of all pages from a particular website can be evaluated to determine whether the website as a whole can be deemed as a suspicious or bad website.

In our other approach to this problem we employ the technique of topic modeling, to identify salient topics in the sentences present as a part of the various healthcare websites. Topics are identified using TagMe and correspond to the titles of articles available in Wikipedia. An analysis of the similarity measures the topics identified is used to decide if the information from candidate website falls under the suspicious or trustworthy category.

In our approach using the Hidden Markov Model we chose to use all the candidate sentences available from a website for building our good and bad sentence models rather relying only on sentences which were identified to contain salient topics. This is due to the fact that some of the sentences that contain characteristics of a trustworthy or suspicious information may not be necessarily be identified to contain a salient topic.

6.3.1 Hidden Markov Model Analysis

6.3.1.1 Sentence Classification

In general, an HMM is used to create a model with hidden states that represent the latent characteristics of the pattern that we are trying to model, which however emit symbols or observations that are visible. In case of sentence modeling, the hidden states would represent the characteristics of the sentence that we are trying to model while the words forming the sentence would represent the visible observations. HMM uses a supervised learning method were a training set is provided to train the model to identify sentences represented by the model. The Baum Welch algorithm is used to train HMM to learn the transition and observation probabilities. Once trained, the HMM can then be used for computing the probability of a sentence belonging to given model using the Forward-Backward algorithm or can be used to predict the possible hidden state sequence that could have generated a given sequence of observations using the Viterbi algorithm.

In our approach we create two separate HMMs: one to model the suspicious sentences and the other to model trustworthy sentences. Once these HMMs are trained, the probability of any new sentence belonging to both the models is determined. The sentence is then classified to belong to a model which has the highest probability value. The following sections detail the construction of the two sentence models using HMM.

Suspicious Sentence Model Following are some of the features present as a part of the sentences from suspicious websites which have been used to build a Hidden Markov model that is representative of such sentences.

1. Most of suspicious websites often contain testimonials or personal experiences of people trying to promote or sell a particular product. Most often the information in such sites presented in a highly subjective manner with first person narratives describing their experiences. Such sentences can be easily identified with the start of sentence having the pronoun I. Example: a. I thought you should know that I have now lost 23 pounds, beyond the 11 pounds when I wrote the testimonial.

2. In addition to first person experiences, some of the suspicious web sites also have second person directives where the author provides instructions for the reader to follow. Again these are with the intent of coercing the user to avail some services or make purchases in order to address their problem. Most of these can be identified by the presence of modal verbs like should, must, need to, have to and ought to that help enforce the directions. Example: a. You must combine the diet pill usage with diet plans and do at least some aerobics for half an hour.

3. In most cases the information presented in the suspicious sites tend to resort to superlatives in order to sell the products or services. These can be identified by the presence of superlative like most, best, worst, etc in the sentences. Examples: a. This is the best method and I am shocked that dermatologists don't recommend it. b. The best thing is that the pill works in the human intestines and therefore it does not have many side effects.

4. With the intent of the suspicious sites in general being to sell some kind of service or product to the end user, we can identify the presence of several commercially related terms appearing as a part of the content. Examples: a. Subscribe today and receive four free ebooks worth $60 . b. You have the same no-risk, money-back offer on this all-time best-selling ebook as you do with all burn the fat products.

5. In addition we would also find occurrences of some suspicious words and content that appear to promise a guaranteed solution with no scientific study or references to back them up. Most of websites that talk about using supplements, hypnosis, holistic or alternative treatments fall under this category. Examples: a. This amazing beverage helps fight your risk of heart attack and heart disease in several other ways. b. It's irrefutable - even neuroscientists are now forced to agree - this powerful formula for getting a stunning body with mind power is fool proof!

6. Finally one other commonly observed pattern among the content from suspicious websites is the presence of several sentences that end with an exclamation mark. Most often these sites tend to convey strong feelings in emphasizing the capabilities of a product or in commanding the user to take an action and hence end with an exclamation. Examples: a. I will definitely buy this tea again! b. Guaranteed results or your money back! c. Stop procrastinating!

States of the Suspicious HMM

So the different features or characteristics identified earlier can be combined to represent the states of a Hidden Markov Model that can be used to model a sentence from a suspicious website. Since a HMM is typically used to model sequential data, the above features can be represented using the following states to model a sentence containing a sequence of words or tokens. Since some of these features can be easily identified if the sentence is annotated with Part-Of-Speech tags, we make use of the Stanford NLP tagger to identify the POS tags for the sentences.

- Personal Pronoun (PRP). This is a part-of-speech tag that can be used to represent the personal pronoun like I and you which can used to identify the first person narratives and second person directives in the sentences
- Modal verbs (MD): Identifies modal verbs like should, must, need to, ought to and have to found in sentences commanding a user to take action
- Superlatives: Presence of superlative adjectives or superlative adverbs identified with the part-of-speech tags JJS and RBS respectively
- Commercial terms: Presence of the following set of commercial terms or keywords. This list was manually extracted based on term frequency analysis on the extracted content from the suspicious websites. Table 6.1 shows the examples of commercial terms.
- Suspicious terms: Presence of the following set of suspicious terms or keywords. This list was manually extracted based on term frequency analysis on the extracted content from the suspicious websites. Table 6.2 shows the examples of some suspicious terms.
- Exclamation: Presence of the exclamatory sign at the end of the sentence.
- Other: The other state is used to represent words that do not fallen under any of the above mentioned states.

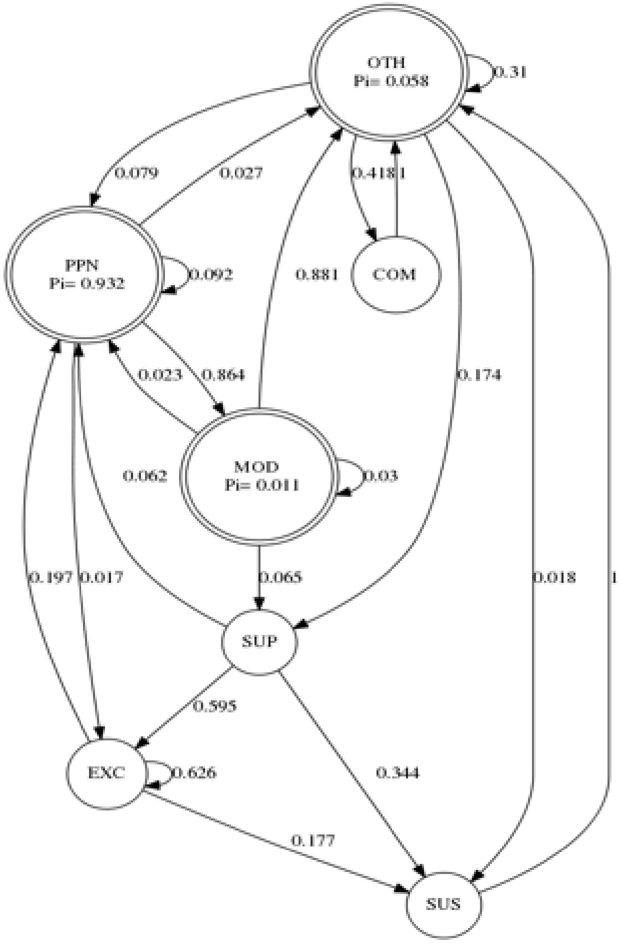


Figure 6.2: State transitions in a trained HMM representing a suspicious sentence model

Table 6.1: List of Commercial terms

money	advertise	purchased	prizes	fees	order
money-back	advertised	customer	ebook	shop	orders
pay	loan	customers	ebooks	shopping	ordering
payment	loans	dollar	discount	world-class	popular
payments	service	dollars	discounts	membership	worldwide
repayment	services	endorsement	cheap	one-time	product
shipping	program	endorsements	expensive	offer	products
free	programs	price	inexpensive	sale	review
buy	purchase	prices	subscription	sales	reviews
buying	purchases	credit	fee	sold	cash
payday	compensated	affiliate	affiliates	bonus	bonuses

Table 6.2: List of Suspicious terms

hypnosis	proven	success	burn
hypnotherapy	guarantee	successful	burning
hypnotherapists	guaranteed	story	flawless
self-hypnosis	trust	stories	result
supplement	trusted	lifesaver	results
supplements	testimonial	formula	amazing
pill	testimonials	formulas	holistic
pills	secret	reliable	session
alternative	secrets	magic	sessions
healing	lifetime	magical	perfect
miracle	miracles	miraculous	

Figure 6.2 displays the states of the suspicious Sentence Model HMM, after being trained on the data from the training set. The probability values displayed on the arcs are the transition probability values. The nodes with a double circle are used to denote the possible starting states for the HMM and have Pi value denoting the probability of starting in that state.

Trustworthy Sentence Model

Following are some of the features present as a part of the sentences from trustworthy sites that have been used to build a HMM that is representative of such sentences.

1. Most the sentences that provide credible information often tend to be expressed in Passive voice. The passive voice can be detected in general by the presence of the

Noun-Verb-Noun format in the sentence where the first noun is typically the Object and the second noun is usually the Subject. So in general identifying the Noun-verb-noun format helps to identify possible credible sentences. Examples: a. CRE infections(N) are caused(V) by a family of germs(N) that are a normal part of a person's healthy digestive system. b. Aneurysms(N) can also be caused (V) by serious infections (N).

2. In case of the passive voice the verb forms that are used tend to be either:

   - a "be" form verb like be, am, is are, was, were, been, being
   - a "have" form verb like have, has, had, having
   - a past tense verb
   - a particle verb

   Examples: a. Infection risk can be reduced by practicing good hygiene, such as washing hands often b. These are found in the blood of patients who have been infected with EBV.

3. The nouns that are used as part of the sentence could be of the following form:

   - Proper nouns that are usually names of things or places.
   - Gerunds, verbs followed by -ing, acting as nouns, eg: Smoking, Cycling, etc.
   - Also nouns other than Proper nouns that can appear in the singular or plural form.

   Examples: a. A bronchoscope is inserted through the nose or mouth into the trachea and lungs. b. Neuroendocrine tumor cells take up the radioactive MIBG and are detected by a scanner. c. Stretching and weight training can also strengthen your body and improve your fitness level. d. Coughing up mucus is often the first sign of COPD.

States of the Trustworthy HMM In order to represent the above characteristics to model a Trustworthy sentence model we define the following states of the HMM:

- Proper Noun (PRN): Identifies nouns that may be names of places, things, etc.
- Be Form Verb (BFV): Identifies verbs like: be, am, is are, was, were, been, being
- Have Form Verb (HFV): Identifies verbs like: have, has, had, having
- Past Tense Verb (PTV): Identifies a verb mentioned in the past tense
- Gerund (GER): Identifies if a word is a gerund
- Participle Verb (PAV): Identifies if a verb is present participle or past participle
- Other Noun (OTN): Identifies nouns other than proper nouns
- Other words (OTH): Identifies all other words that do not fit into the above categories

Figure 6.3 displays the states of the Trustworthy Sentence Model HMM, after being trained on the data from the training set. The probability values displayed on the arcs are the transition probability values. The nodes with a double circle are used to denote the possible starting states for the HMM and have Pi value denoting the probability of starting in that state.

Sentence Classification into Trustworthy and Suspicious models For each sentence the probability of it belonging to either of the models is computed and the sentence is classified to belong to the model which has the highest probability.

6.3.1.2 Page classification

Once the sentences have been classified using the HMM based classifiers, their probabilities and frequency counts are used to further classify if the page containing those sentences is trustworthy or not. A Support Vector Machine based classifier is explored to make this prediction.

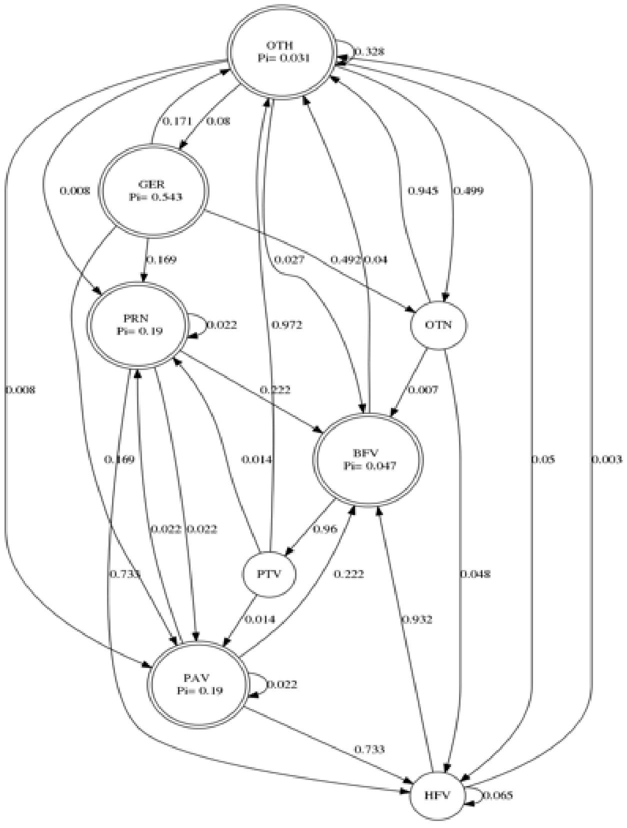


Figure 6.3: State transitions in a trained HMM representing a suspicious sentence model

Page Level Features For every web page based on the HMM based sentence classification we compute the counts for total number of sentences, number of trustworthy sentences, number of suspicious sentences and number of neutral sentences. In order to obtain the page level features, for each sentence in a page we compute:

$$D = Probabilityofsentencebeingtrustworthy?Probabilityofsentencebeingsuspicious \tag{6.1}$$

The value of the above difference, D, is then used to update the counts of a histogram whose values range from -1.0 to +1.0. Each bucket in the histogram has an interval which covers the range of 1.0E-10. The counts are then normalized by the total number of sentences in the page. In all there are 23 buckets in the histogram.

6.3.2 Topic Analysis

6.3.2.1 Term-distribution Analysis

Traditional *topic modeling* approaches are based on "bag-of-words" model. Therefore the frequency of the terms is an important factor for the topic analysis. To analyze the term distributions of the health-related web pages, firstly we construct a medical-related term dictionary from MedlinePlus [1] built by NIH. We adopt the health topic categories from MedlinePlus which is classified into thirty categories. Each category contains medical terms or general healthcare terms. Some terms appear redundantly in several categories, however, we collect the terms whereas they are in. In addition, we allow the *exact match* for topics for multiple words such as 'breast cancer'. Table 6.3 shows the thirty categories and the number of terms in the category. The total number of terms is 2243.

In order to find the term distributions of authoritative and suspicious sites, we chose 'www.cdc.gov' and 'top10weightcontrol.com' as sample sites. Figure 6.4 and Figure 6.5 show the term distributions of both sites respectively. As shown in the figures, 'cdc.gov' matches many terms from MedlinePlus, especially Pregnancy and Reproduction (157 terms), Trans-

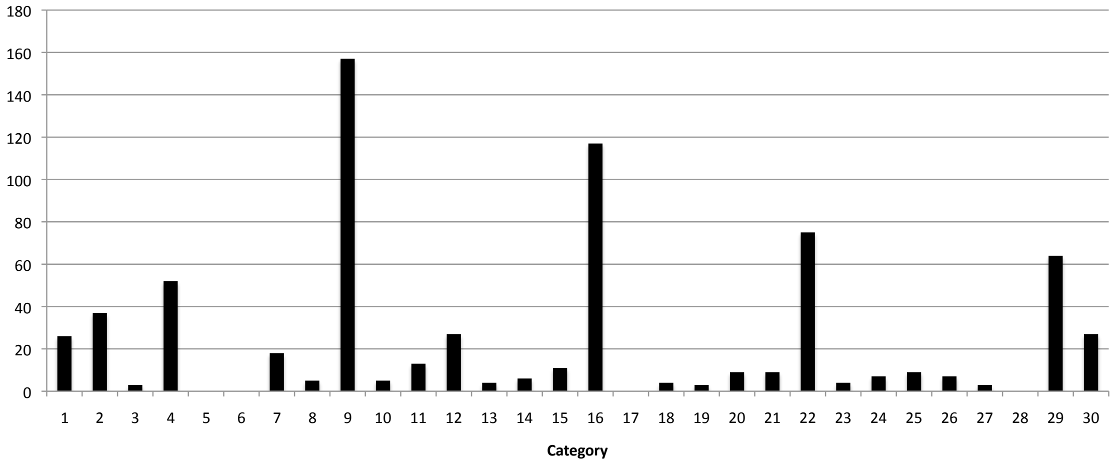


Figure 6.4: Term distribution of **www.cdc.gov**

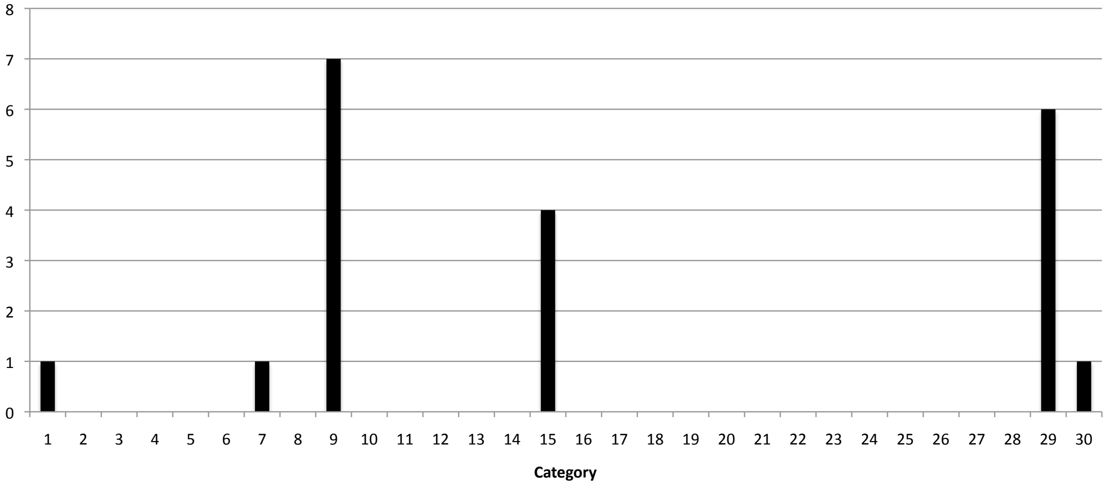


Figure 6.5: Term distribution of **top10weightcontrol.com**

plantation and Donation (117 terms), and Disasters (75 terms). However, most terms in thirty categories appeared in the 'cdc.gov' web site. Compared to 'cdc.gov', the suspicious site 'top10weightcontrol.com' only matched few terms among the categories such as Pregnancy and Reproduction (7 terms), Symptoms (4 times), and Social/Family Issues (6 terms). Based on the term distributions, we expect that topics identified from MedlinePlus can be used to distinguish the trustworthy sites and suspicious sites. However, this term dictionary cannot cover all general health related terms or similar concepts in the text, we need a more systematic method to identify the topics.

Table 6.3: The number of terms in the category

Health Topics	Category	The number of terms
Disorders and Conditions	1. Cancers	150
	2. Diabetes Mellitus	170
	3. Genetics/Birth Defects	15
	4. Infections	30
	5. Injuries and Wounds	27
	6. Mental Health and Behavior	27
	7. Metabolic Problems	50
	8. Poisoning, Toxicology, Environmental Health	13
	9. Pregnancy and Reproduction	120
	10. Substance Abuse Problems	102
Diagnosis and Therapy	11. Complementary and Alternative Therapies	57
	12. Diagnostic Tests	253
	13. Drug Therapy	128
	14. Surgery and Rehabilitation	55
	15. Symptoms	82
	16. Transplantation and Donation	50
Demographic Groups	17. Children and Teenagers	29
	18. Men	56
	19. Population Groups	32
	20. Seniors	98
	21. Women	46
Health and Wellness	22. Disasters	111
	23. Fitness and Exercise	51
	24. Food and Nutrition	49
	25. Health System	59
	26. Personal Health Issues	69
	27. Safety Issues	90
	28. Sexual Health Issues	22
	29. Social/Family Issues	61
	30. Wellness and Lifestyle	141

6.3.2.2 Topic Identification

In order to perform topic analysis, web pages from a manually selected set of 20 trustworthy and suspicious sites were gathered to form our reference sites. Plain text sentences were extracted from these pages and TAGME was used to identify semantic topics in these sentences. As described in Section 2.4.1.1, TAGME provides the goodness value of a topic and its corresponding topic word. We used a certain goodness threshold for choosing meaningful topics among all the topics in a sentence obtained from TAGME.

6.3.2.3 Page-Level Similarity

The similarity between pages is measured after identifying the semantic topics for each page. Since each page contains representative topics, we can compare the similarities with each other. For calculating page-level similarity, we use the most popular set-similarity method called *Jaccard similarity* and it can be obtained by:

$$PageSim(x_i, y_i) = \frac{Total\ number\ of\ same\ topics}{Total\ number\ of\ topics} \tag{6.2}$$

where x_i and y_i are pages from each website. Based on this, we can calculate the page-level similarity score between pages.

6.3.2.4 Site-Level Similarity

Next, in order to get the site-level similarity, we adopted *Group Linkage* problem. [43] Group Linkage problem is used to identify the similarity between two groups which have different number of elements. The fundamental idea of the group linkage is based on *Maximum Bipartite Matching(MBM)* problem. [61] The definition of MBM similarity is defined as below:

Let A and B be two sets, $A = a_1, a_2, ..., a_m$ and $B = b_1, b_2, ..., b_n$. The Maximum Bipartite Matching Similarity, *MBM_Sim*, is defined as:

$$MBM\_Sim(A, B) = \frac{\sum_{(a_i,b_j)\in M} sim(a_i, b_j)}{m + n - M} \tag{6.3}$$

where $sim(a_i, b_j) \geq \rho$ is the similarity of two elements in the two groups A and B. M is the number of maximum weight matching in the bipartite graph. In our approach, $sim(a_i, b_j)$ would be the page-level similarity above ρ. The threshold ρ is to remove the pages having very low similarity scores and can be decided heuristically. Using MBM_Sim, we get the similarity score between two web sites.

Finally, in order to evaluate the trustworthiness of unknown site X, we defined the *ContentSim(X)* as below:

$$\begin{aligned} ContentSim(X) = {} & [Sum\ of\ Top5\ Site - Level - Similarity\ of\ Trustworthy\ Sites] \\ & - [Sum\ of\ Top5\ Site - Level - Similarity\ of\ Suspicious\ Sites] \end{aligned} \tag{6.4}$$

The content similarity of site X, ContentSim(X), is the difference between the summation of top 5 site-level-similarity from trustworthy group and the summation of top 5 site-level-similarity from suspicious sites. If the value of *ContentSim* score is positive, we consider that site X contains trustworthy information. If not, it contains suspicious information.

6.4 Integration

We developed a new algorithm to integrate two components of HealthTrust, Structure and Content-based analysis for getting the final HealthTrust scores; the first one is the trust ranking scores ranged between 0 and 1, showing how much the website is relatively trustworthy based on the link structure; and the second one is similarity scores ranging from positive to negative scores, showing that how much the contents of website has similar semantics compared to the positive and negative seed sets. Thus we can measure the *trustworthiness from the structure* and the *semantics* from contents by the two methods. Overall, our new integrating algorithm should combine those two orthogonal concepts effectively. Thus we expect the integrated algorithm to distinguish each node's characteristics even though a node has high trust ranking scores but low semantic scores, the combined scores should represent the different characteristics of both analyses.

6.4.1 Overview of the algorithm

Structure-based Analysis is solely based on the link structure of the web graph. However, we use the TwoWay-TrustRank algorithm to propagate the trustworthiness of a website considering its positive and negative factors. Also, Content-based Analysis is performed using the semantics of the sentences in a document. Therefore the characteristics of the two methods are orthogonal. A challenging problem is that how we can combine them without losing their orthogonal features. In order to integrate effectively, we developed a new iterative algorithm that performs what it needs to be as we mentioned above.

6.4.2 Integration of Structure Analysis and Content Analysis (ISACA)

The overall algorithm procedure is shown in Figure 6.6. The inputs for the ISACA algorithm are two scores from Structure-based analysis and Content-based Analysis, two sets from positive and negative seed sets and the list of all websites in the web graph. First, we run the TrustRank algorithm to find out the first rank of all sites with the positive and negative sets. For decay factor for TrustRank, we set the $d = 1$ for the positive seed set and $d = -1$ for the negative seed set. Then we sort the new TrustRank list by descending order and find fixed number of websites, *StepSize*=k, for top k sites into the positive and negative seed sets. For positive seed set, we select top sites and for negative seed set, we select lowest ranking sites.

Then based on the updated seed sets after running TwoWay-TrustRank, we run Content-based algorithm with identified topics for all sites using TAGME. Content-based algorithm also provides similarity scores for all sites to the reference seed sets. The rank list is sorted by descending order based on the similarity. Like the previous step, the positive and negative seed sets are updated by the new content-similarity scores. However, we want to integrate the structure-based approach and content-based approach, so we find the intersection of both sets and generate the integrated positive and negative seed sets for next iterations.

Next, we need to consider when the iterative routines should be stopped. In order to identify the stop condition, we compare the two final ranked lists from the Structure-based analysis and those of Content-based analysis. We check the distance for each site between the two lists using RMSD error. For example, if site A is ranked 5 in the TwoWay-TrustRank (TR) and ranked 9 in the Content Similarity (CR), then the distance between two values are $(9-5)^2$ and then we can calculate the RMSD distance error between two lists by the equation (6.5).

$$RMSD = \sqrt{\frac{\sum_{t=1}^{n}(TR(X_t) - CR(X_t))^2}{N}} \tag{6.5}$$

We expect the value of RMSD to decrease as the algorithm iterates. Therefore if RMSD is less than the threshold , we don't need to go further for next iteration step. However, if the seed sets are very different from each other, we then go to the step 2 to run the TrustRank again and so on. Therefore the algorithm runs iteratively until the seed sets are very similar to each other. We set a threshold to make the algorithm converge after a certain number of iterations.

Step Size The step size (StepSize) for updating the seed sets can be decided heuristically depending on the data set size. Therefore the step size should be reasonably chosen; $(StepSize) \leq (Total\ number\ of\ sites)$. If the StepSize is too bug then the seed sets would increase too fast. And if it is too small, then the runtime of the algorithm would be very slow in case the data set is really big. However, we need to consider that if we include the seed set from the best and worst cases in the lists all the time, then the seed set is hardly changed so we might always get similar seed set list after several iteration.

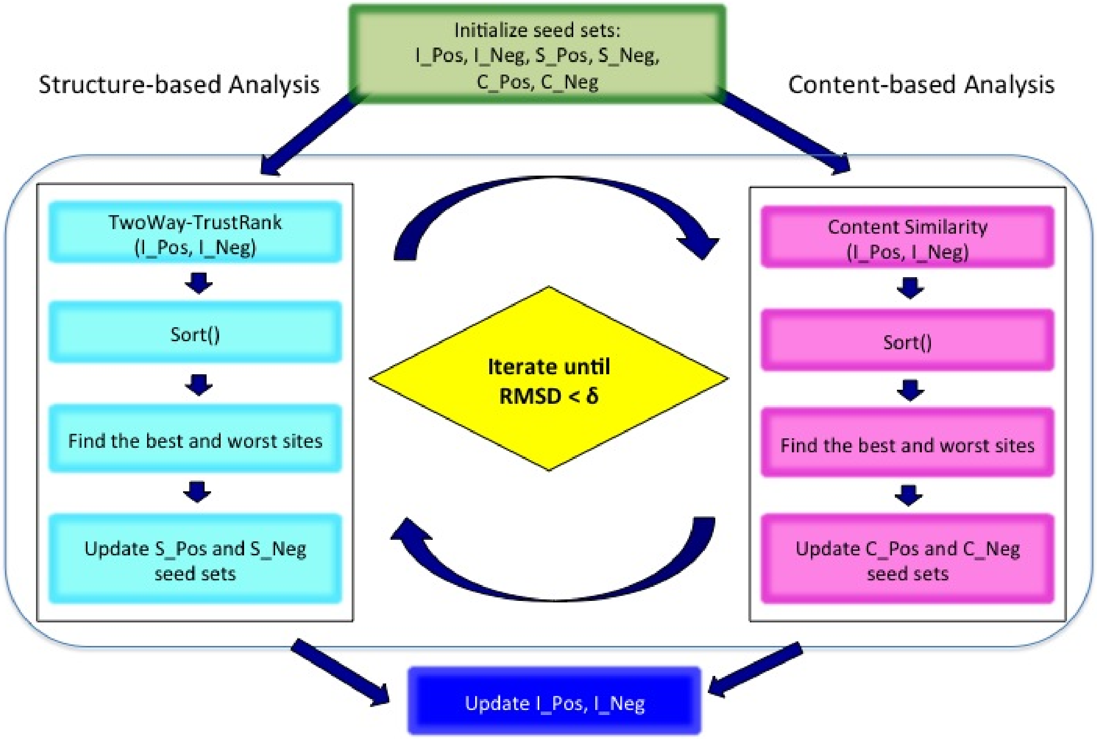


Figure 6.6: Integration algorithm

Chapter 7

Experiments and Results

7.1 Data Set

In order to construct a healthcare related Web graph, we use the crawled data set as we described in Chapter 3. We gather healthcare webpages by crawling the Web. Figure7.1 shows the process of constructing a Web graph in the diagram. We construct a Web graph by analyzing the degree of incoming and outgoing links of the pages. In our Web graph, the nodes can be webpages or websites. Since the data set is huge, the Web graph is very complicated.

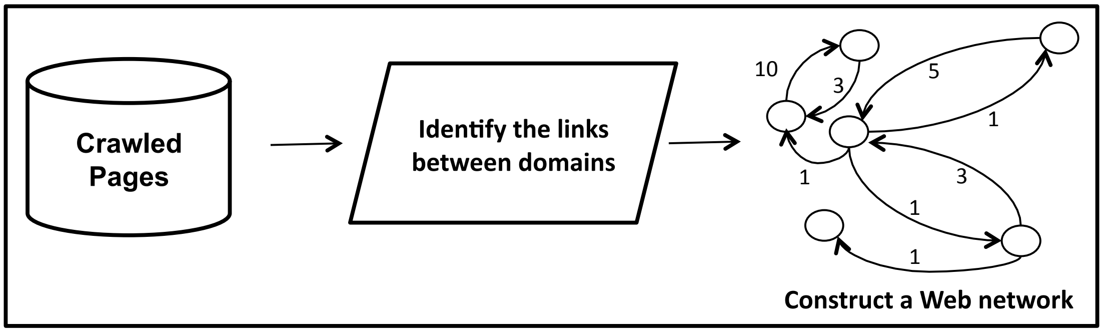


Figure 7.1: Data Preparation

7.1.1 HMM Data Set

7.1.1.1 Manually annotated Seed Data Set

In order to train and test the HMMs a Seed Data set consisting of a small number of sentences were manually annotated to form the train and test sets for both the trustworthy and suspicious sentence models.

Suspicious Sentences Training Set For the suspicious sentences training set, 150 sentences, 15 from each of the following 10 suspicious sites were manually annotated and used for training the HMM:

- amazing-green-tea
- apple-cider-vinegar-benefits
- burnthefat
- carallumaburnreviews
- dietprescriptions-rx
- eco-diet
- fatvanish
- hypnosisnetwork
- fatburningfurnace
- healthynewage

Suspicious Sentences Validation Set All sentences extracted from the following suspicious sites are used as a part of the validation data set to evaluate the classification accuracy for suspicious sentences predicted using the HMM classifiers:

- amazing-green-tea

- apple-cider-vinegar-benefits
- best-colon-cleanse
- burnthefat
- calorie-count
- calorieking
- dieting4weightloss
- eco-diet
- fatvanish
- Herbalife
- hypnosisnetwork
- weightloss-diet-facts

Trustworthy Sentences Validation Set All sentences extracted from the following trustworthy sites are used as a part of the validation data set to evaluate the classification accuracy for trustworthy sentences predicted using the HMM classifiers:

- cancer
- cdc
- drugs
- flu
- nih

7.1.1.2 Real Data Set

After the validation, the classification accuracy of the HMM based classifiers are evaluated on the Real Data set. ***Suspicious Sentences Real Data Set*** All sentences extracted from the following suspicious sites are used as a part of the Real Data Set to evaluate the real classification accuracy for suspicious sentences predicted using the HMM classifiers:

- blogtalkradio
- comcblog
- devinalexander
- directselling411
- discovergoodnutrition
- flite
- goodhousekeeping
- losethebellyfatnow
- ocregister
- planetarynutrition
- plentyofhealth
- premadeniches
- wholefoodsmarket
- widgetbox
- zendesk

Trustworthy Sentences Real Data Set All sentences extracted from the following trustworthy sites are used as a part of the Real Data Set to evaluate the real classification accuracy for trustworthy sentences predicted using the HMM classifiers:

- ChooseMyPlate
- clinicaltrials
- dana-farber
- diabetes
- drugabuse
- foodsafety
- hhs
- kidshealth
- letsmove
- mayoclinic
- nemours
- nutrition
- usa
- webmd
- womenshealth

7.2 Evaluation

7.2.1 HMM evaluation

7.2.1.1 Sentence Classification

Sentence Classification with the Seed Data Set The following Table ?? provides the results of the classification on the Seed Data Set consisting of both the suspicious and Trustworthy train and test data sets.

Data set	Total sentences	Number of classified as suspicious	Number of classified as trustworthy	Number of classified as neutral	Classification Accuracy
Suspicious Train Set	150	149	1	0	99.33%
Suspicious Test Set	120	0	150	0	100%
Trustworthy Train Set	150	0	150	0	100%
TrustworthyTest Set	150	39	111	0	74%

Table 7.1: Sentence Classification for manually annotated Seed Data Sets

Sentence Classification with the Validation Data Set The following Table 7.2 provides the results of the classification on the Validation Data Set consisting of sentences from the suspicious sites.

The following Table 7.3 provides the results of the classification on the Validation Data Set consisting of sentences from the trustworthy sites.

Sentence Classification with the Real Data Set The following Table ?? provides the results of the classification on the Real Data Set consisting of sentences from the suspicious sites.

Website	Total sentences	Number of classified as trustworthy	Number of classified as suspicious	Number of classified as neutral	Classification Accuracy of Suspicious sentences	Baseline Website classification based on sentence classification
Herbalife	1993	1029	959	5	48.12%	Trustworthy
amazing-green-tea	14130	5199	8889	42	62.91%	Suspicious
apple-cider-vinegar-benefits	57093	11647	45327	119	79.39%	Suspicious
best-colon-cleanse	6440	3350	3089	1	47.97%	Trustworthy
burnthefat	9233	2487	6717	29	72.75%	Suspicious
calorie-count	325	100	223	2	68.62%	Suspicious
calorieking	1711	534	918	259	53.65%	Suspicious
dieting4weightloss	3316	1434	1878	4	56.63%	Suspicious
eco-diet	6745	2604	4135	6	61.30%	Suspicious
fatvanish	279473	23766	253560	2147	90.73%	Suspicious
hypnosisnetwork	32057	11883	15631	4543	48.76%	Suspicious
weightloss-diet-facts	4968	1909	3059	0	61.57%	Suspicious

Table 7.2: Sentence Classification on Validation Data Set for Suspicious Sites

Website	Total sentences	Number of classified as trustworthy	Number of classified as suspicious	Number of classified as neutral	Classification Accuracy of Suspicious sentences	Baseline Website classification based on sentence classification
cancer	8592	7647	729	216	89%	Trustworthy
cdc	3430	2223	1160	47	64.81%	Trustworthy
drugs	13773	7219	6534	20	52.41%	Trustworthy
flu	1509	1008	498	3	66.8%	Trustworthy
nih	7167	5302	1702	163	73.98%	Trustworthy

Table 7.3: Sentence Classification on Validation Data Set for Trustworthy Sites

7.2.1.2 Page Classification

From the Validation data sets for the trustworthy and suspicious websites, sentence level classification is carried out for all sentences in all the pages belonging to those sites. The page level features as mentioned before are extracted for all the pages. In all 2329 instances of suspicious page features are extracted from the suspicious Validation data set and 1140 instances of trustworthy page features are extracted from the trustworthy Validation data set. The extracted page level features from the Validation data set are then used from the training and testing data set for the SVM classifier that would be used for page classification.

SVM Cross-validation The extracted feature values from the Validation data set are then scaled to normalize the values of the features. A 10-fold cross validation is then carried out to identified the best parameters for the SVM. Parameter values of $C = 128$ and $g = 0.5$ were identified as the best parameters after the cross-validation giving an accuracy of 92.15% Those values are then used to actually train the full training data set to form the learnt SVM model. The classification accuracy of the trained SVM on the training data set was 95.96%.

Page classification on Validation data set The following Table 7.5 lists the classification accuracy for the SVM based page classifier for the pages available for training and

Website	Total sentences	Number of classified as trustworthy	Number of classified as suspicious	Number of classified as neutral	Classification Accuracy of Suspicious sentences	Baseline Website classification based on sentence classification
blogtalkradio	7852	3908	3637	307	46.32%	Suspicious
comcblog	9872	4810	4933	129	49.97%	Suspicious
devinalexander	1950	402	1541	7	79.03%	Suspicious
directselling411	4315	2046	2260	2443	52.38%	Suspicious
discovergoodnutrition	3474	1031	1702	0	70.32%	Suspicious
flite	3742	1826	1913	3	51.12%	Suspicious
goodhousekeeping	5315	1626	3530	159	66.42%	Suspicious
losethebellyfatnow	2235	704	1531	0	68.50%	Suspicious
ocregister	7514	3570	3579	365	47.63%	Suspicious
planetarynutrition	2516	1252	1264	0	50.24%	Suspicious
lentyofhealth	8383	4667	3715	1	44.32%	Trustworthy
premadeniches	5205	1376	3829	0	73.56%	Suspicious
wholefoodsmarket	295	236	56	3	18.98%	Trustworthy
widgetbox	5375	1018	4357	0	81.06%	Suspicious
zendesk	2369	652	1697	20	71.63%	Suspicious

Table 7.4: Sentence Classification on Real Data Set for Trustworthy Sites

Website	Total pages	Number of classified as trustworthy	Number of classified as suspicious	Classification Accuracy of Suspicious pages	Website classification based on page classification
Herbalife	199	19	180	90.45%	Suspicious
amazing-green-tea	171	9	162	94.74%	Suspicious
apple-cider-vinegar-benefits	194	9	185	95.36%	Suspicious
best-colon-cleanse	176	1	175	99.43%	Suspicious
burnthefat	81	4	77	95.06%	Suspicious
calorie-count	200	1	199	99.5%	Suspicious
calorieking	192	10	182	94.79%	Suspicious
dieting4weightloss	396	0	396	100.0%	Suspicious
eco-diet	402	2	400	99.50%	Suspicious
fatvanish	240	26	214	89.17%	Suspicious
hypnosisnetwork	802	2	800	99.75%	Suspicious
weightloss-diet-facts	392	12	380	96.94%	Suspicious

Table 7.5: Page Classification on Validation Data Set for Suspicious Sites

testing in the Validation set for the suspicious sites.

The following Table 7.6 lists the classification accuracy for the SVM based page classifier for the pages available for training and testing in the Validation set for the trustworthy sites.

Page classification on Real Data Set Similarly for the real data set, in all 3295 instances of suspicious page features are extracted from the real suspicious data set and 2882 instances of trustworthy page features are extracted from the trustworthy real data set. The following Table 7.7 lists the classification accuracy for the SVM based page classifier for the pages available in the Real data set for the suspicious sites.

Website	Total pages	Number of classified as trustworthy	Number of classified as suspicious	Classification Accuracy of Suspicious pages	Website classification based on page classification
cancer	256	249	7	97.27%	Trustworthy
cdc	252	241	11	95.63%	Trustworthy
drugs	200	184	16	92.0%	Trustworthy
flu	98	95	3	96.94%	Trustworthy
nih	334	305	29	91.32%	Trustworthy

Table 7.6: Page Classification on Validation Data Set for Trustworthy Sites

Website	Total pages	Number of classified as trustworthy	Number of classified as suspicious	Classification Accuracy of Suspicious pages	Website classification based on page classification
blogtalkradio	200	64	136	68.0%	Suspicious
comcblog	309	119	190	61.49%	Suspicious
devinalexander	168	25	143	85.12%	Suspicious
directselling411	203	102	101	49.75%	Trustworthy
discovergoodnutrition	192	38	154	80.21%	Suspicious
flite	233	104	129	55.36%	Suspicious
goodhousekeeping	196	8	188	95.92%	Suspicious
losethebellyfatnow	154	42	112	72.73%	Suspicious
ocregister	189	97	92	48.68%	Trustworthy
planetarynutrition	200	176	24	12.0%	Trustworthy
plentyofhealth	201	31	170	84.58%	Trustworthy
premadeniches	175	21	154	88.0%	Suspicious
wholefoodsmarket	199	5	194	97.49%	Trustworthy
widgetbox	478	70	408	85.36%	Suspicious
zendesk	198	24	174	87.88%	Suspicious

Table 7.7: Page Classification on Real Data Set for Suspicious Sites

Website	Total pages	Number of classified as trustworthy	Number of classified as suspicious	Classification Accuracy of Suspicious pages	Website classification based on page classification
ChooseMyPlate	200	97	103	48.5%	Suspicious
clinicaltrials	200	196	4	98.0%	Trustworthy
dana-farber	198	147	51	74.24%	Trustworthy
diabetes	200	97	103	48.5%	Suspicious
drugabuse	192	38	154	80.21%	Suspicious
foodsafety	50	33	17	66.0%	Trustworthy
hhs	310	213	97	68.71%	Trustworthy
kidshealth	187	83	104	44.39%	Suspicious
letsmove	127	59	68	46.46%	Suspicious
mayoclinic	294	112	182	38.10%	Trustworthy
nemours	200	162	38	81.0%	Trustworthy
nutrition	69	22	47	31.88%	Suspicious
usa	103	50	53	48.54%	Suspicious
webmd	198	27	171	13.64%	Suspicious
womenshealth	166	89	77	53.61%	Trustworthy

Table 7.8: Page Classification on Real Data Set for Trustworthy Sites

The following Table 7.8 lists the classification accuracy for the SVM based page classifier for the pages available in the Real data set for the trustworthy sites. On average the overall classification accuracy on the real test data set was 61.62%.

7.2.1.3 Website Classification

Classification Accuracy of Baseline Sentence Classification vs SVM Page Classification Table 7.9 below compares the website classification accuracy of the baseline sentence classification approach to the SVM based page classification approach.

Approach	Data Set	Total Numbe of web sites	Number of classified as trustworthy	Number of classified as suspicious	Classification Accuracy
Baseline	Real Data for Suspicious web sites	15	3	12	80%
Baseline	Real Data for Trustworthy web sites	15	10	5	66.67%
SVM Page Classification	Real Data for Suspicious web sites	15	3	12	80%
SVM Page Classification	Real Data for Trustworthy web sites	15	6	9	40%

Table 7.9: Website Classification Baseline vs SVM Page classification

7.3 Results

7.3.1 Data Set

We tested our algorithm to real data set collected by the focused crawler. However, we found that only 387 sites have the outgoing links to other sites. All other nodes are dangling points or there are no connected sites. Therefore we ignore other sites in this experiments. The real data set statics are shown in the Table ??. We gathered 63,894 files from the 387 sites and there are 59,702 non-empty files kong them. The total number of sentences is 1,873,486. We preprocessed the raw html files removing all HTML tags and Javascript languages. Furthermore, HMMs need natural language grammar, we made all sentences capitalized at the beginning of the sentences which make it possible to understand sentences.

7.3.2 Topic Analysis Results

We used TAGME web application to get the semantic topics from the sentences including short-text in a webpage. For example, we gathered all the pages from *www.nih.gov* and one

Data set	Total number of Files	Total number of Non-empty Files	Total number of Empty Files	Total number of Sentences
Total	63,894	59,702	4,192	1,873,486
Average	176.65	165.03	11.62	5194.75

Table 7.10: Summary of Real Data Sets

of the webpages contains mostly disease-related words or general words instead of showing specific topics or product names shown in Table **??**. The first column is a topic for the words and the second column is a related category in Wikipedia.

However, TAGME provides a threshold how much we want to cut off the range of topics. If we want to strict range of topics from a specific category, we can set the goodness of the topic discovered which is called rho. The higher the threshold is, the mpre accurate topics we can find. Therefore, according to the user's criteria, we can set the value differently.

7.3.3 HMM Alnaysys Results

We tested our HMM analysis method using the real data set. As our evaluation shows the HMM analysis works poor compared to Topic analysis. In real data set, it shows similar results to the evaluation results. Most government sites are classified well in our data set. However, many trustworthy sites sponsored by organizations or commercial industries are not well classified. Because those web sites uses general terms or sentences instead of scientific sentences. Since we strictly modeled the trustworthy HMM, many commercial sites are classified as suspicious. Table 7.12 shows the government sites in our real data set and classified correctly. The examples for correctly classified suspicious sites are shown in Table 7.13. However, Table 7.14 listed examples of wrong classified sites as shown below.

Vaccine:Vaccine
Simian immunodeficiency virus
Monkey AIDS
HIV:AIDS Virus
Virus:Virus
Immune Attack:immune attack
Understanding:Understanding
Testosterone:Testosterone
Human:Men
Causality:effects
Body:body
Estrogen:estrogen
Brain tumor:brain tumors
Brain:Brain
Integrated circuit:Circuit
Visual system:Visual
Developmental biology:Development
Mouse:mouse,Research:study,Human eye:eyes
Therapy:treating
Amblyopia:amblyopia
Amblyopia:lazy eye
Bacteria:Bacteria
Neuron:Nerve Cells
Etiology:Cause
Pain:Pain
Stimulation:stimulate
Sensory neuron:sensory neurons
Neuron:neurons
Inflammation:inflammation
Insight:insights
Therapy:treatments

Table 7.11: Examples of topics discovered by TAGME

Website	Total Pages	Trustworthy Pages	Suspicious Pages	Unclassified Pages	HMM Classification	Manual Classification
childwelfare.gov	291	274	17	0	Trustworthy	Trustworthy
cdc.gov	290	245	45	0	Trustworthy	Trustworthy
ahrq.gov	301	261	40	0	Trustworthy	Trustworthy
dhhs.gov	1	1	0	0	Trustworthy	Trustworthy
cms.gov	303	270	33	0	Trustworthy	Trustworthy
bls.gov	304	272	32	0	Trustworthy	Trustworthy
ca.gov	118	80	38	0	Trustworthy	Trustworthy
cancer.gov	225	187	38	0	Trustworthy	Trustworthy
clinicaltrials.gov	299	223	76	0	Trustworthy	Trustworthy
childrenwithdiabetes.com	301	152	149	0	Trustworthy	Trustworthy
allaboutvision.com	289	224	65	0	Trustworthy	Trustworthy

Table 7.12: Examples of Trustworthy Sites by HMM

Website	Total Pages	Trustworthy Pages	Suspicious Pages	Unclassified Pages	HMM Classification	Manual Classification
doctorline.com	272	114	158	0	Suspicious	Suspicious
dietriffic.com	302	22	280	0	Suspicious	Suspicious
central.com	69	24	45	0	Suspicious	Suspicious
coolnurse.com	1	0	1	0	Suspicious	Suspicious
brettterpstra.com	1	0	1	0	Suspicious	Suspicious
drgreene.com	282	10	272	0	Suspicious	Suspicious
cancercompass.com	323	138	185	0	Suspicious	Suspicious

Table 7.13: Examples of Suspicious Sites by HMM

Website	Total Pages	Trustworthy Pages	Suspicious Pages	Unclassified Pages	HMM Classification	Manual Classification
diabetesmonitor.com	301	123	178	0	Suspicious	Trustworthy
cancercenter.com	314	41	273	0	Trustworthy	Suspicious
best-home-remedies.com	303	280	23	0	Trustworthy	Suspicious
diabetes.org	329	20	309	0	Suspicious	Trustworthy
aafp.org	301	134	167	0	Suspicious	Trustworthy
amia.org	304	122	182	0	Suspicious	Trustworthy

Table 7.14: Examples of Wrong Classified Sites

7.3.4 Integration Results

We ran the ISACA algorithm using the real data set and found the updated positive seed sets and negative seed sets from both analysis. Among them, we selected 20 sites inside the positive and negative seed sets shown in Table **??** and Table **??**. We started with small number of seeds and expand the seed set after several iterations. The results showed that most sites in positive seed sets are correctly grouped and negative seed set also contains mostly suspicious sites.

Furthermore, we found that after running the TwoWay-TrustRank algorithm and Content Similarity separately, they have the same sites with 80% and 52% for positive and negative seed set respectively. The results are shown in Table 7.15 and 7.16.

7.3.4.1 RMSD Results

The graph 7.2 shows that the RMSD error rate is decreased after a few iterations. At the beginning, RMSD is increased and then started decreasing at some point as we expected. It means that after several iterations, the ranking lists from both structure- and content-based analysis agree on choosing the trustworthiness sites cooperating the characteristics of each other.

7.3.4.2 Integrated Seed Set Results

Since we take the integrated seed sets for both Structure and Content-based analysis in each iterations after updating the seed sets, **??**figseedsets

7.3.4.3 HealthTrust Ranking Results

We find the HealthTrust scores for each site by calculating the average of two scores from Structure- and Content-based analysis. Table 7.17 shows the top 15 websites after performing the ISACA algorithm and all sites within Top 15 is trustworthy. It outperforms link analysis

No.	Website
1	healthsquare.com
2	cdc.gov
3	inserm.fr
4	healthscout.com
5	msfocus.org
6	nextgen.com
7	ahaf.org
8	flu.gov
9	biochemj.org
10	medicinenet.com
11	hhs.gov
12	dhhs.gov
13	healthcare-ny.com
14	molinahealthcare.com
15	ccfa.org
16	healthcare411.org
17	healthonnet.org
18	hopkinsmedicine.org
19	nih.gov
20	cancer.gov

Table 7.15: Examples of Positive Seeds

No.	Website
1	expasy.ch
2	naturalsolutionsmag.com
3	momsmedicinechest.com
4	alternativedr.com
5	actagainstaids.org
6	fathersfirstyear.com
7	natural-homeremedies.com
8	central.com
9	medpagetoday.com
10	e-health-europe.com
11	mendosa.com
12	libertybella.com
13	fioricetnow.com
14	modernhealthcare.com
15	demandbase.com
16	homeremedypro.com
17	alsa.org
18	mymigraineconnection.com
19	acceleratedcure.org
20	mood247.com

Table 7.16: Examples of Negative Seeds

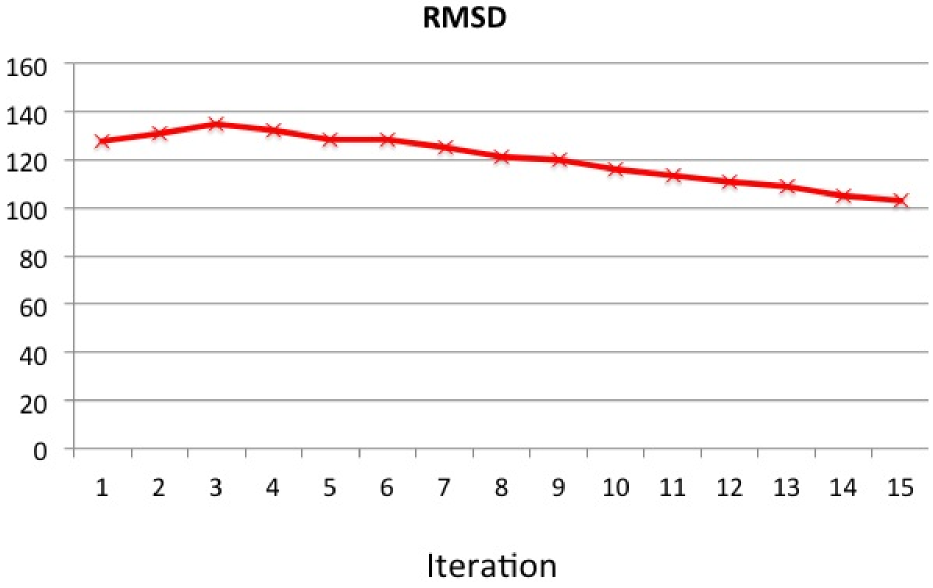


Figure 7.2: RMSD Results

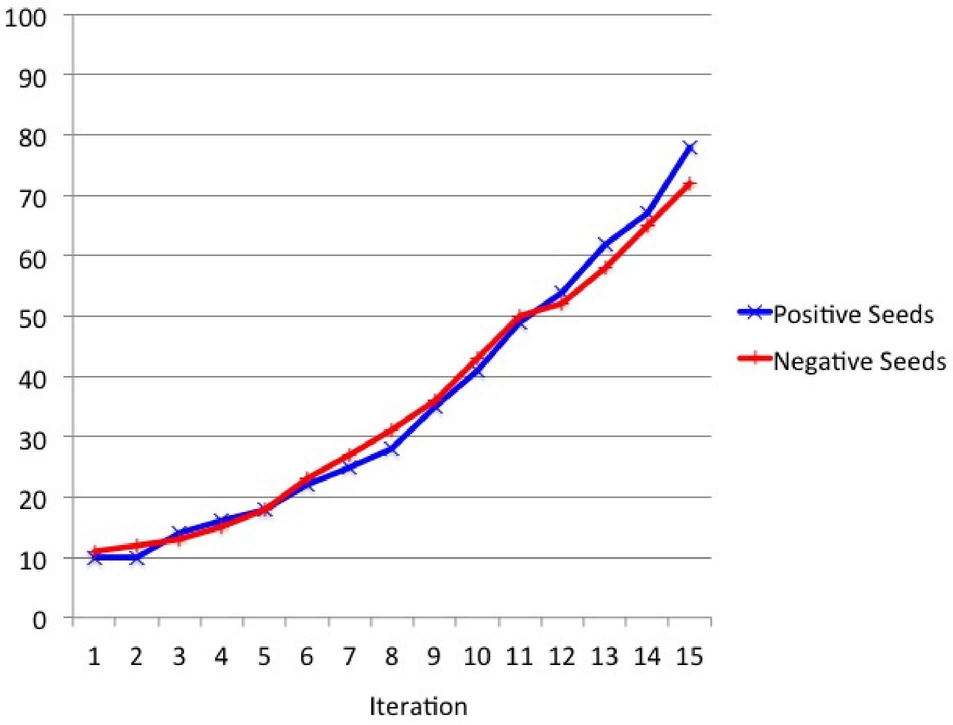


Figure 7.3: Number of Integrated Seed Sets after Each Iteration

No.	Website
1	dhhs.gov
2	medlineplus.gov
3	pelvicpainnewyork.com
4	familydoctor.org
5	jdrfcapitol.org
6	medscape.com
7	smokefree.gov
8	helpingamericayouth.gov
9	healthcare411.org
10	ncpublichealth.com
11	unc.edu
12	clevelandclinic.org
13	texaspain.org
14	ivfspecialists.com
15	otcsafety.org

Table 7.17: HealthTrust Results: Top15

approaches. We conclude that our ISACA algorithm compromises well the link analysis and content analysis

Chapter 8

Conclusion

Healthcare Informatics is a promising field to utilize the flood of healthcare related data through adoption of information technologies. Currently, a large amount of data are produced in this field along with the web technologies such as social networking services, real-time web technologies such as wikis, blogs, and RSS (Really Simple Syndication). However, not all health information provided online is trustworthy. Even though many experts are involved in publishing trusted information, people can hardly determine the credibility of the information easily. Most search engines have the ability to control spam pages, but cannot determine the trustworthiness of the page yet. Hence, identifying the credible information on the complicated web society is a challenging problem.

8.1 Summary of Research

In this dissertation, I did a thorough study on a large volume of online healthcare information collected by a focused crawler. With the focus on the providers of such information, I have tried to answer questions such as who they are, how they are distributed, and how they are related. I have measured the network structural features, analyzed the graph topology, and studied the nodes and the links distributed over top level domains. I have used link analysis approaches, PageRank, HITS, and TrustRank to study the authoritativeness of websites based on graph topology, and found that traditional approaches gave mixed results of credible, suspicious, and irrelevant sites. With two user studies, I showed that commercial search engine results for health-related queries were far from satisfactory; many untrustworthy or highly suspicious sites were returned. Meanwhile it is not easy for users to distinguish trustworthy sites from the mixed results. A reliable mechanism to automatically determine the trustworthiness of online healthcare information is highly desired.

First of all, I proposed a new system named **HealthTrust** to provide the overall assessment system which automatically assesses the trustworthiness of healthcare information over the Internet. In HealthTrust system, "Structure-based Analysis" is performed in order to use the link-structure of the web graph, which is a traditional method to assess the rankings of the web pages. However, link analysis has its own drawback which is not considering the trustworthiness of the contents. In order to overcome the limitations, we need a method to start propagating the trustworthiness through the web graph. We used "TwoWay-TrustRank" which takes into account the positive and negative factors. However, structure-based analysis fundamentally has the limitations, since it only takes into account the link structure rather than the content of the web pages.

Therefore, for next step, I developed "Content-based Analysis" which is based on topic modeling and machine learning techniques. We followed two methods for analysis: (1) *Topic discovery*: short-text tagging application called TAGME is used to identify salient topics in the sentences available in the healthcare websites. An analysis of the similarity measures among the topics identified is used to decide if the information from candidate website falls under the suspicious or trustworthy category. (2) *HMM analysis*: Hidden Markov Models are applied to model trustworthy and suspicious sentences using an annotated training set.

Finally, in order to integrate the two approaches, I used an iterative algorithm that integrates the credibility assessments from structure-based and content-based methods. In the iterations, strongly positive and strongly negative results from the structure-based approach will be used as "additional seeds" in the content-based approach, and vise versa. The iterative approach further counteracts the problem of limited seeds as well as the sparseness of the document space.

I believe that HealthTrust will improve the existing way of finding useful information related to healthcare for hundreds of millions of the Internet users. However, the trend of the Web has been rapidly changing to focus on Social Networks (SNs) environment such as Facebook, Twitter. The consumers tend to heavily rely on the contents or comments of SNs as well. In addition, the characteristics of the SNs are very different from other scientific, authoritative websites. Therefore it needs more careful observations and approaches to assess the trustworthiness of the SN information. We consider the SNs as a huge Web graph where the users correspond to nodes, and hyperlinks or *like* correspond to links between nodes. The comments or messages correspond to the text information. We can apply our HealthTrust system to the SN graph to identify the most important factors that affect users.

8.2 Future Work

Despite the fact that the HealthTrust and existing methods trying to assess the trustworthiness of the information related to healthcare, we are still far away from obtaining accurate assessment due to the characteristics of the natural language and the lack of verifying systems of the healthcare information. Since it is controversial that the content of healthcare websites is true or not, the reference sets are not always true. Moreover, I only explored a small part of the web embracing certain topics in this dissertation. In fact, the Internet itself is a tremendously huge graph and it is really difficult to cover the whole scale of it by a focused crawler to get the expected coverage. There is still a great need for developing more efficient algorithms to deal with the big web data and controversial contents problem.

Therefore, the problem description is, how we can accurately assess the contents. We need to take into account the contents of web pages from all various websites than its link structure. In computer science perspective, the problem is to include how to deal with the *big-data* efficiently. Nowadays, dealing with *big-data* becomes a hot topic due to rapidly growing social networking sites as well as general web sites. Healthcare informatics is one of the challenging problem.

For future studies, firstly, I would like to elaborate my content-based analysis by modeling sophisticate algorithm. My goal is to assess the contents with accuracy and speed. As part of my research, I would like to address how we can assess the contents of web pages in real time without waiting for the results. Secondly, the integration algorithm should be improved to identify more accurate trustworthy and suspicious groups

References

[1] Medline plus, 2012.

[2] A user's guide to finding and evaluating health information on the web @ONLINE, 2012.

[3] K. Adelhard and O. Obst. Evaluation of medical Internet sites. *Methods of Information in Medicine*, 39:75–79, 1999.

[4] P. Berkhin. A survey on pagerank computing. *Internet Mathematics*, 2(1):73–120, 2005.

[5] T. Berners-Lee and D. Connolly. Hypertext markup language-2.0. Technical report, RFC 1866, November, 1995.

[6] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology (TOIT)*, 5(1):231–297, 2005.

[7] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[8] C. Burroughs and F. Wood. Measuring the difference: Guide to planning and evaluating health information outreach. 2000.

[9] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.

[10] R. Cline and K. Haynes. Consumer health information seeking on the internet: the state of the art. *Health education research*, 16(6):671–692, 2001.

[11] A. Coulter, V. Entwistle, and D. Gilbert. Sharing decisions with patients: is the information good enough? *Bmj*, 318(7179):318–322, 1999.

[12] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716, 2007.

[13] J. Culver, F. Gerr, and H. Frumkin. Medical information on the internet. *Journal of General Internal Medicine*, 12(8):466–470, 1997.

[14] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proceedings of the third conference on Applied natural language processing*, pages 133–140. Association for Computational Linguistics, 1992.

[15] D. Davies and D. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979.

[16] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[17] J. Diaz, R. Griffith, J. Ng, S. Reinert, P. Friedmann, and A. Moulton. Patients' use of the internet for medical information. *Journal of general internal medicine*, 17(3):180–185, 2002.

[18] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon. Pagerank, hits and a unified framework for link analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 353–354. ACM, 2002.

[19] L. Egghe. Mathematical relations between impact factors and average number of citations. *Information processing & management*, 24(5):567–576, 1988.

[20] L. Egghe and R. Rousseau. Introduction to informetrics: Quantitative methods in library, documentation and information science. 1990.

[21] G. Eysenbach, T. Diepgen, J. Gray, M. Bonati, P. Impicciatore, C. Pandolfini, and S. Arunachalam. Towards quality management of medical information on the internet: evaluation, labelling, and filtering of informationhallmarks for quality of information-quality on the internetassuring quality and relevance of internet information in the real world. *Bmj*, 317(7171):1496–1502, 1998.

[22] G. Eysenbach and A. Jadad. Evidence-based patient choice and consumer health informatics in the internet age. *Journal of medical Internet research*, 3(2), 2001.

[23] G. Eysenbach and C. Köhler. How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. *Bmj*, 324(7337):573–577, 2002.

[24] G. Eysenbach, J. Powell, O. Kuss, and E. Sa. Empirical studies assessing the quality of health information for consumers on the world wide web. *JAMA: The Journal of the American Medical Association*, 287(20):2691–2700, 2002.

[25] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM, 2010.

[26] W. Francis and H. Kucera. Frequency analysis of english usage. 1982.

[27] B. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[28] C. Friedman, G. Hripcsak, et al. Natural language processing and its future in medicine. *Acad Med*, 74(8):890–5, 1999.

[29] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.

[30] J. Giménez and L. Marquez. Svmtool: A general pos tagger generator based on support vector machines. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*. Citeseer, 2004.

[31] L. Greenberg, U. D'Andrea, and U. Lorence. Setting the public agenda for online health search. *URAC*, 2003.

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

[33] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[34] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466. ACM, 2009.

[35] J. Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

[36] A. Langville and C. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.

[37] A. Leithner, W. Maurer-Ertl, M. Glehr, J. Friesenbichler, K. Leithner, and R. Windhager. Wikipedia and osteosarcoma: a trustworthy patients' information? *Journal of the American Medical Informatics Association*, 17(4):373–374, 2010.

[38] J. e. a. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA, 1967.

[39] M. Marcus, M. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[40] F. Meric, E. Bernstam, N. Mirza, K. Hunt, F. Ames, M. Ross, H. Kuerer, R. Pollock, M. Musen, and S. Singletary. Breast cancer on the world wide web: cross sectional survey of quality of information and popularity of websites. *Bmj*, 324(7337):577–581, 2002.

[41] M. Mézard. Where are the exemplars? *Science*, 315(5814):949–951, 2007.

[42] A. Ng, A. Zheng, and M. Jordan. Stable algorithms for link analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 258–266. ACM, 2001.

[43] B.-W. On, N. Koudas, D. Lee, and D. Srivastava. Group linkage. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 496–505. IEEE, 2007.

[44] F. Osareh. Bibliometrics, citation analysis and co-citation analysis: A review of literature i. *Libri*, 46(3):149–158, 1996.

[45] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[46] G. Peterson, P. Aslani, and K. Williams. How do consumers search for and appraise information on medicines on the internet? a qualitative study using focus groups. *Journal of Medical Internet Research*, 5(4), 2003.

[47] L. Rabiner and B. Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

[48] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[49] A. Ratnaparkhi et al. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, PA, 1996.

[50] S. Rieh. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53(2):145–161, 2001.

[51] T. Robinson, K. Patrick, T. Eng, D. Gustafson, et al. An evidence-based approach to interactive health communication. *JAMA: the journal of the American Medical Association*, 280(14):1264–1269, 1998.

[52] N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. J. Tick. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, 1(2):142–160, 1994.

[53] P. Spyns. Natural language processing. *Methods of information in medicine*, 35(4):285–301, 1996.

[54] P. Stavri, D. Freeman, and C. Burroughs. Perception of quality and trustworthiness of internet resources by personal health information seekers. In *AMIA Annual Symposium Proceedings*, volume 2003, page 629. American Medical Informatics Association, 2003.

[55] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[56] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424, 2006.

[57] K. Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North*

American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 173–180. Association for Computational Linguistics, 2003.

[58] K. Toutanova and C. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics, 2000.

[59] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web*, pages 585–594. ACM, 2006.

[60] E. R. Weitzman, E. Cole, L. Kaci, and K. D. Mandl. Social but safe? quality and safety of diabetes-related online social networks. *Journal of the American Medical Informatics Association*, 18(3):292–297, 2011.

[61] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Englewood Cliffs, 2001.

[62] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492. ACM, 2012.

[63] J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

[64] G. Zhou and J. Su. Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480. Association for Computational Linguistics, 2002.