Automated Shortlived Website Detection

A study and evaluative prototype

Subramaniam Ramasubramanian

Kongens Lyngby 2015
COMPUTE-M.Sc-2015-303

COMPUTE-M.Sc-2015-303

ISSN: 0909-3192
Technical University of Denmark
DTU COMPUTE - Department of Applied Mathematics and Computer Science
Richard Petersens Plads
Building 324
DK-2800 Kongens Lyngby
Denmark

Phone +(45) 45 25 33 51
Phone +(45) 45 88 26 73

compute@compute.dtu.dk
www.compute.dtu.dk

Preface

This thesis was prepared under the guidance of Professor Christian D. Jensen at the department of Informatics at the Technical University of Denmark and Professor Markus Hidell from the School of Information and Communication Technology at KTH Royal Institute of Technology in fulfillment of the requirements for acquiring an M.Sc degree in Security and Mobile Computing.

The work presented in this thesis was supported by DBI who provided support in terms of requirements and business domain specific knowledge.

Lyngby, 26-June-2015

R.Subramani

Subramaniam Ramasubramanian

Table of Contents

Page

List of Figures

List of Acronyms

DBI	-	The Danish Institute of Fire and Security Technology
DNS	-	Domain Name System
API	-	Application Programming Interface
HTML	-	HyperText Markup Language
HTTP	-	HyperText Transfer Protocol
JSON	-	JavaScript Object Notation
UI	-	User Interface
DB	-	DataBase
PL/SQL	-	Procedural Language/Structured Query Language
CLOB	-	Character Large Object
ER Diagram	-	Entity Relation Diagram
IDE	-	Integrated Development Environment
IP	-	Internet Protocol
VPN	-	Virtual Private Network
TOR	-	The Onion Router
DHCP	-	Dynamic Host Configuration Protocol
ISP	-	Internet Service Provider
MAC	-	Media Access Control
CAPTCHA	-	Completely Automated Public Turing test to tell Computers and Humans Apart
DOS	-	Denial Of Service

Summary

Counterfeit pharmaceutical products are a big threat to the society not only because of the monetary losses incurred by ineffective drugs but also because of the adverse effects they cause to consumers.

It is becoming increasingly more common for these products to find their way to the customer through websites that are marketed in the open Internet.

We work with key stakeholders from research and industry to develop approaches to solve the three key problems of discovering new websites that sell these products, automatically identifying websites that sell these products and classify them into meaningful groups of websites that can be analysed together.

The project also produced a working prototype tool that is used in order to test these approaches identified and documents/analyse the results produced by the tool.

It was observed that the use of user dictionary based mechanisms to discover, identify and rank these websites demonstrated the capability to produce exceptionally high quality results.

Acknowledgements

Firstly, I would like to thank my mom and dad who were so supportive of their only son's idea of giving up a plush job and going to a country far far away to study something obscure and unintelligible to the common man. This project is as much a product of your hard work indirectly as it is of mine directly.

I would also like to thank Michael Lund and DBI for providing not only excellent support and domain knowledge for the project but also for being extremely warm, courteous and friendly to me personally. The constant feedback, quick replies and keen interest that were showed helped me stay motivated and deliver strong results.

Thanks to Christian, Markus and Luke for giving me a free hand and the gentle nudges in the right direction whenever I needed them.

Special thanks to Keahi, and the others from NordSecMob and all 16 of my past/present roommates for putting up with my childish antics and helping me remain focused during my work hours.

I would also like to thank Keld, and all my teammates in Danske Bank for being flexible and accommodating, fitting their schedule around mine throughout this semester!

And lastly, I would like to thank the sun, the moon and the stars for being in the right place and letting all good things happen the way they did.

CHAPTER 1

Background and Introduction

"In God we trust; all others must bring data."

*(*W. Edwards Deming*)*

1.1 Introduction

This chapter provides a general outline of the problem that the thesis aims to address and its relevance in the real world. It clearly establishes the motivation for research on the topic and how to achieve them.

We then proceed to state the objectives of the project. The report also provides an overview of the requirements for a tool that is an ideally expected by-product of this thesis as described by its primary stakeholders in this chapter under the requirements section.

We then highlight the major approaches to solving the stated problem and the corresponding reasoning for having chosen them for testing and experimentation during the course of the project.

A brief overview of the results obtained is also presented at the end.

1.2 Background

Counterfeit pharmaceutical products are a big threat to the society not only because of the monetary losses incurred by ineffective drugs but also because of the adverse effects they cause to consumers owing to improper preparation or dosage [6].

It is becoming increasingly common for such drugs to be produced, manufactured or acquired illegally in one part of the world and then marketed and shipped throughout the world through the use of websites and the Internet [13].

Because of the open nature of the Internet and its penetration, it is extremely easy for a person on one side of the world to use it and acquire these pharmaceutical products smuggled and shipped from the other!

Besides causing huge losses for the pharmaceutical companies that develop these drugs, their impact on the society is also grave. People without proper medical knowledge and training are enticed into using these products without taking the proper precautions or being aware of possible side effects. Studies have shown how grievously the health of regular users of human growth hormones for non medical purposes can be affected through side effects [17].

There have also been numerous studies on how counterfeit pharmaceutical products which either did not go through the right process of preparation or lacked standard quality assurance practices are recking havoc in many countries.

Some of these products are either completely useless to the consumer in which case the sellers are just scamsters, but in some cases they cause sever side effects because of lack of knowledge of how to use these drugs or in the most grievous cases cause harmful effects to the consumers because they contain completely unrelated drugs in the packaging [4].

The lack of quality control, transparency and accountability in these operations causes much of the problem.

1.2.1 Social and Ethical impact

The social and ethical impact of the problem at hand is enormous. Public health and safety are of paramount importance to every country.

The value of a human life is immeasurable and should always be protected. Uninformed recreational body builders predominantly young adults (around 30 years of age) form the bulk of the audience that these websites target. The younger demographic of the victims and long term nature of the side effects that these drugs can have in improper dosage make a deadly mix that can leave families and societies devastated [2].

Other key stakeholders that can be impacted by this project are pharmaceutical companies that manufacture these drugs. Studies and research show that it takes in some cases more than 2 billion dollars to develop a market approved drug [1]. There is a lot of effort, research and resources that go into the development of these drugs and any activity that can undermine the funds that these organizations receive can be viewed as a direct threat to their growth and development [5]. If these organizations are not able to reap the complete rewards of their hard work, it can lead to slacked development of drugs that can combat new diseases in the future. Moreover it is ethically and legally wrong for someone to steal the fruits of other peoples hard work.

The value that this project can contribute in terms of helping both of these sectors and certain key stakeholders makes it worth investing time and energy into.

1.2.2 Key Stakeholders

Up until now, pharmaceutical companies have typically hired private investigators to find and monitor, from the open web, online shops that sell relevant products and used help from international law enforcement agencies to bring them to justice. Private investigators have typically had to manually search the Internet and track leads from user forums and search providers with variations of search terms. Following this, they gather and analyse vast amounts of data as evidence manually. Though the existing manual process is very reliable in identifying leads and gathering evidence, there is a constant risk of newer websites coming up under the investigative radar. Also manually analysing websites is a very slow and cumbersome process. It also becomes virtually impossible to identify similarities and classify websites that have been tracked and monitored by different investigators in a team even though it is for a single company.

Companies typically use the information provided by these private firms in order to improve the safety and security processes within their own supply chain and manufacturing departments. National law enforcement agencies are also tightly pressed for time to go out into the Internet and find crime where it happens as they are bombarded with many other issues.

DBI is one such organization that primarily deals with fire safety but is also venturing into brand protection for their customers. A key part of brand protection is to check the sale of counterfeit products of clients in the market. Their primary stake in the project is to provide assistance in gathering data that can be used for testing and investigative expertise in order to develop a tool that helps ease the load of its investigators and augment their capabilities.

1.3 Objective

Though counterfeit products are being sold widely in a large number of places, we restrict our scope to counterfeit products marketed online in this study.

The focus is to develop strategies to discover potential websites, from the open web, that could be used in this deadly supply chain from various sources of input; then come up with mechanisms to identify and enrich patterns that help rank or validate the potential danger they pose.

The secondary objective is the development of a working prototype which utilizes the approaches tested above and is equipped with these features that aid investigators bring law enforcement closer to shutting such websites down effectively.

Criminal investigation is as much about patient and careful documentation to build evidence as it is about discovering leads. Hence it is vital for us to be able to document, track and monitor these websites once they are identified. Furthermore, given the sheer volume of the instances of active websites that are selling counterfeit pharmaceutical products we can see that mere documentation of these results would simply result in an overwhelming amount of data that would not be useful if not sorted by relevance.

The relatively little investment in this section of cyber crime, specifically the discovery and prosecution of criminal sale of counterfeit products online, from law enforcement agencies makes it even more important for us to invent new ways of identifying the most active websites that cause the most tangible damage. This would enable their elimination and help address the issue from its root upwards. For example, it is easier to focus on shutting down one website which is potentially supplying ten other websites if we can identify and prove their role using evidence. This would curb the influence of many players in the market much more efficiently with limited effort from law enforcement.

This also makes it important for the tool to be as automatic, user friendly and use as little of the investigators time in tuning and maintaining the tool and the various techniques it uses in order to function accurately.

Another key requirement to note is that, even though the pharmaceutical project is a low hanging fruit in the orchard, the techniques and strategies identified through the course of this project are applicable, with very little modifications, to a vast array of other generalized website identification, classification and discovery problems. It is also desirable that the prototype developed in this project can be reused with little effort to evaluate its potential in other unrelated fields as well. Hence it is essential that the concepts used in the development of the tool are mathematically sound and can be used to produce stable results in a wide spectrum of similar issues.

It is also important to note that with the ever increasing focus on Big Data applications in various industries, it might also be possible in the future for us to mine the vast amounts of data that the tool would collect in order to arrive at conclusions that we cannot possibly foresee given our current insignificantly little understanding of the problem in the larger context of things. Thus flexibility ground up becomes a necessity throughout the development of the tool and its strategies.

1.4 Requirements Analysis

In order to better understand the domain of online criminal investigation and the mechanisms used by private investigators to discover, identify, store and monitor websites, a shadowing exercise was taken up where a real investigators day was closely followed during his daily activities. This helped produce a preliminary set of requirements which were a reflection of the abstract question of how to make things easier for the investigator to find and prosecute online criminals who host websites that sell counterfeit products.

These were then discussed with DBI who are the key stakeholders for the project, it has been identified that the following key features are relevant to the project from an investigative perspective.

The features expected out of the tool by DBI in this regard can be broadly classified into three main branches. The first being, the ability of the tool to take a new website and identify if it conforms to patterns seen in the general database of websites that the investigators are interested in. The second being, the ability of the tool to look into the database and come up with classifications and sub patterns within the collected database of websites automatically. The third being, the ability of the tool to go out into the open web and find more websites that the investigator might be interested in, based on the database.

Hence it made logical sense for us to classify the tools capabilities into three engines that were closely modeled to fit these high level requirements accurately. These are the Identification, Classification and Discovery engines respectively for each feature of the tool.

The further subsections address the requirements based on their classification into one of the three categories. The other chapters in the report also use these terms consistently in order to explain the working and evaluation of the tool.

Note: This is merely an ideal preliminary feature set description from DBI and is different from the actual tool developed during the project and even the modes of how each operation is performed by the tool to reach the same goal.

1.4.1 Identification Engine

The identification engine works with user defined and machine learned patterns to help provide a ranking for newly discovered websites on a well-defined scale. The user is able to create, categorize and organize patterns groups and the corresponding patterns. Once such patterns are defined, the tool matches these patterns to the database of websites that is available to then produce a rank for each unique entry based on parameters such as the content, the amount of match, the accuracy of each match and the number of matched patterns.

- The user defined patterns set the precedence for the direction in which a search is filtered but the machine learned patterns are a means to keep the patterns fresh and sharp. They are to be built as a database based on user feedback of identified websites. For instance, if the user reviews a website that was flagged and "dislikes" it, then the patterns that were used to match that particular web site are weighed down accordingly. If on the other hand a website is "liked" by the user, then the patterns used to match it are weighed more.
- Also this process is used to identify the similarities between the "liked" and "disliked" pages and further generate more patterns which can be weighed to keep the tools identification as sharp as the initial input and even improve them on the long run with minimal manual labor.
- Classification engine: A sub feature of the identification engine, in that this module needs to be able to identify websites that are similar to each other, perhaps even developed by the same programmer in a huge list of flagged websites. This helps law enforcement agencies prioritize websites as causing more damage to the market based on factors such as spheres of influence (within Europe, within Denmark etc.), volume of transactions and bring in real world importance of acting on bringing a particular website down.
- More advanced features in the identification engine would be to be able to categorize and order images, discover which of them are new, mine for meta-data within them and be able to provide a complete picture of what is being analyzed.

1.4.2 Discovery Engine

The discovery engine is the part of the project that revolves around providing mechanisms to feed the identification engine. A few areas that were discussed preliminarily below.

- A web-crawler to crawl through known forums of discussion, known websites selling counterfeit products and other known sources to discover potential new entries as soon as they appear. The key is to appear as close to a normal user as possible through randomized access, randomization in scheduling of jobs and to overcome weak captcha images through image recognition and stronger ones through manual user input. Special emphasis also needs to be placed on providing an anonymous environment from which the tool can access the websites to leave behind little trace.

- A honey-pot email address through which phishing related inputs can be drawn and analyzed to produce accurate and more diverse sources of web sites.
- Analysis of redirect requests from known sellers of counterfeit products. Here, all requests to a blacklisted website are checked to identify where the users original request originated from and if it was from another website, then the identification engine attempts to identify if that is also a potential candidate. A similar strategy applied to child pornography websites, has been proven successful and may be applied to this problem as well.
- Analysis of DNS request patterns to spot where traffic flows appear similar and spot potential similarities between different destinations to help identify spurious websites that sell counterfeit products.

1.4.3 Other Non-functional requirements

Other non-functional requirements that are of significance are as follows

- A means to constantly identify and learn more discovery techniques needs to be identified. Learning techniques can be applied to the discovery module as well to ensure high quality of the output at all times.
- A good coupling between the discovery engine and the identification engine would result in a system that learns by itself from the user input
- A good system is also expected to provide reporting and customization features along with diverse search functionality to quickly filter based on any parameter, on the identified websites.
- Ability to limit the resources being used by the tool, keeping the long running jobs fail safe and robust with graceful failure is also of paramount importance.
- A simple but elegant user interface to access these features with minimal human intervention is also a must.
- To overcome the captcha problem, the possibility of a browser plug-in tool to capture details during a normal users browsing session was also discussed and needs to be studied to find if it is a necessity.
- Ability to create multiple isolated jobs with different databases and learning needs is also desirable.

1.5 Overview of results

A prototype of the tool with all three broad functions described under the Discovery, Identification and Classification Engine was built and verified to produce valuable and accurate results.

The approach of using the vocabulary of websites in both the discovery and identification parts of the website proved highly accurate in most cases.

Visualization of distance measures in terms of multiple parameters for a large dataset of websites proved to be a very useful way to classify websites with little manual effort.

Some of the key non-functional requirements described in the earlier sections of the chapter were incorporated into the tool and the means to achieve the rest are carefully documented as part of the later chapters of this report.

CHAPTER 2

Project Plan

2.1 Introduction

This chapter describes in detail the project plan, the set of tasks to be completed and time-lines for the various deliverables. A discussion of the methodology used in the project to ensure scientific accuracy is also described.

Lastly, an outline of how the rest of the report is structured is presented to make understanding of the subject clearer.

2.2 Project Plan

The project plan gives structure to the project by describing the tasks and timelines with the methodology to be used in order to successfully achieve the goals of the projects.

Based on the requirements specified above, the following set of tasks is to be completed as part of the project. The project is highly experimentation oriented for results to test ideas as there is limited existing research in the area. Also since the focus is on producing a final working prototype, the plan is to iterate cyclically through conceptualize, build, experiment and verify results for each of the two major deliverable aspects of the tool. Another reason to go with the iterative model is that the identification engine and the discovery engine become stronger as each of the other components becomes more efficient. Thus it would be prudent to focus simultaneously on both aspects of the project to use this dependency.

It is agreed that weekly deliverables with an improved version of the tool created every week with emphasis on all aspects of the project as per the requirements to control the project progress under each category and also offer flexibility of thought process and requirements at every stage of the project life-cycle in an "agile" fashion. Taking up the agile methodology ensures that we remain flexible, highly responsive to changes and developments from within or outside the project team and deliver a high quality product [7].

Requirements for the weekly deliverables are considered "frozen" (unchangeable) for the duration of the sprint. Constant discussions with DBI would be taken up to ensure that the projects goals are always aligned and the sprint deliveries are meaningful.

In order to give a measure of how much time would be spent on the various sections of the project on the overall timescale, a rough "number of weeks estimate" for each task is provided but it is very likely that it is not a single monolithic period of time.

2.3 Tasks and Time-line

We attempt to provide an overview of the list of high level tasks to be completed and the time-line for each task as non contiguous number of weeks worth effort. Task dependencies are in the order of listing under each category with each category being reasonably independent.

- Development of Discovery Engine (7 weeks)
  - Study existing methods of discovery
  - Study existing identified malicious web pages
  - Identify new sources of discovery
  - Create a database of meta data information to be gathered by the tool
  - Test source of discovery through implementation in the tool
  - Automate collection of meta data information for discovered web pages
  - Automate manual methods of discovery
  - Develop methods that facilitate the tool to learn new discovery modes
- Development of Identification Engine (7 weeks)
  - Study existing identified malicious web pages for patterns
  - Identify mechanisms to fingerprint the web pages
  - Research alternate sources of identification
  - Research ranking/hashing algorithms to group similar web pages
  - Develop efficient ways of storing/classifying user defined and generated patterns
- Non-functional requirement analysis (3 weeks)
  - Research methods of economization for tools traffic
  - Quantify the need for anonymization in the given context
  - Define and conform to performance metrics
  - Build crash recovery and efficient logging
  - Simple and efficient UI
- Produce a report for the thesis and overall documentation for the tool (5 weeks)

2.4 Evaluation Criteria

A project plan without well defined expectations on output would quickly lose purpose. Hence we provide an evaluation plan that is abstract enough to remain relevant to a dynamic development process and fluctuating requirements but precise in validating if the end goals are met by the project.

1. The tool efficiently identifies web sites with desired malicious content.
2. The tool classifies and groups web sites with a high degree of success.
3. The tool discovers web sites from the open web based on built up database.
4. The tool conforms to performance and anonymization characteristics defined.
5. The tool is robust and handles failures gracefully.
6. The tool works with minimal user intervention and utilizes self-learning techniques.

2.4.1 Practical Notes

To monitor progress and direction for the report, a weekly updated jar file with the latest working version of the software and a description document of what new updates are available will be published in the cloudforge repository Luke has created.

A biweekly update of the methods being researched for implementation and the plan for the upcoming week will also be published in the repository as a separate document.

2.5 Method Description

This chapter describes the research method that will be used by the project to ensure scientific nature of the process involved in delivering the end result and re-usability in other scientific literature and work.

There have been a number of research papers in the area of automated crawling and classification of websites through the identification of patterns in the research. Most of these have been experimentation so those hypotheses that are described earlier could be verified during the course of the project and be proposed as a reasonable solution.

Since our task at hand is focused not only on identifying potentially powerful self-learning mechanisms to discover, identify and classify websites but also on building a working prototype, we follow an experimentation approach where various hypothesis are proposed and evaluated by practically collecting the information needed and analysis them for results using the prototype tool.

The seed list of websites that we will use to gather data is to be provided by DBI. These are assumed to contain data that we would like to discover more of using the end product and verified manually by the investigative team manually at DBI.

The objective is to build the prototype tool as three different but interconnected sub tools each with its specific goal as mentioned below.

- Build a discovery engine based off this seed list and other potential sources to gather more data.
- Build an identification engine that can, given a database of websites that we are interested in, spot other interesting websites from the open web and provide a ranking.
- Build a classification engine that can identify similarities between websites and group them into meaningful classes which may have material sourced from each other.

For each of these sub tools to be built, we propose a hypothesis of a technique that could potentially yield results, build the sub tool, evaluate the results based on manual verification of ground truth with the help of the investigators at DBI.

The reason that the evaluation is to be based on manual verification is that there is no other way for us to obtain ground truth in scenarios such as these where real criminals are running websites that might be interconnected.

The further chapters of this report are also based on this same broad classification (Identification, Discovery and Classification Engines) that is rooted in the various modules of the tool that is to be built. Under each of the chapters, we present the background and the significance of each engine in greater detail based on the data collected from the investigators at DBI, a hypothesis on an approach to solving each of these problems based on observations of investigator behaviour and an evaluation of the results on each hypothesis.

Chapter 3

Discovery Engine

3.1 Introduction

This chapter deals with the detailed description of the features classified under the Discovery Engine part of the tool, its purpose and how the tool handles the feature in its implementation.

A clear hypothesis is stated based on expectations of how the tool is built and what output is expected to the effect of the said hypothesis

We then proceed to document the results and arrive at a conclusion of weather the hypothesis was proved to be true or false.

3.2 Background and Purpose

This part of the project revolves around the creation and maintenance of a healthy database of websites that could contain data of potential interest to investigators in any area. That is, a database of websites that are involved in the sale of counterfeit and illegal prescription-only pharmaceutical products.

The most important leads in this regard are the forums such as those available in reddit where the topics of relevance to us are being discussed online. Some of the forums document explicitly the most active websites selling pharmaceutical products based on user ratings for these websites from their user base. Most well-kept forums have a reasonably updated list of websites in their pages. This can be used as a prime target for web crawling to get high quality results from.

Another intuitive approach here is to search for relevant keywords using common search API providers such as Google and Bing in order to obtain a set of results. Once these results are obtained, it is fairly easy for us to download, maintain and crawl these results in order to document and analyze the content of these websites. But it is highly desirable for the tool to require as minimal human intervention as possible because the keywords that yield good results from search engines keep changing with time for any industry that the investigators deal with.

For instance, the drug being marketed the most in the market now, could be replaced by a newer version or one from a different vendor with a different name within the span of a few months. Additionally, common slang words used by these websites to relate to their customer base could be different depending on the region, time period or target audiences' age group. This would make the existing keywords return increasingly stale results and eventually end up producing completely irrelevant results if not maintained periodically through manual intervention to keep patterns up to date.

We also need to ensure that the web crawler does not explode out into every website it fetches as this could create a huge database with a lot of irrelevant content. Hence only websites which have been flagged as relevant by an investigator will be crawled to any given depth.

It is also important for the Discovery engine to catalog information that is relevant to an investigator in terms of tangible evidence that can be used to build arguments at court against criminals.

3.3 Hypothesis

- If we can automatically produce a set of top “hot” keywords periodically from the seed database we have and query them using the search APIs along with any specific user defined key words, then we can minimize user intervention and keep patterns relevant based on our growing database.
- Crawling webpages to variable depth based on only flagged websites produces data that is predominantly relevant and does not omit important information from being cataloged.

These methods of discovery are hypothesized based on interactions with the investigative team from DBI. An abstraction of the actions that the investigators performed during their tasks was used to formulate the principles that we use to find websites of interest.

3.4 The Tool

This section deals with the description of the tool created for the features described under the Discovery Engine and a description of how each task is achieved.

The three main tasks that the Discovery Engine needs to perform are, gathering web content for any given website, gathering the metadata for each of those webpages and finally mine the website for relevant keywords to search based on.

3.4.1 Gathering Web Content

The first task at hand is to store websites given the link to the HTML from any source. Preliminary use of simple HTML retrieval using HTTP sockets revealed that this method had a few flaws. This technique did not allow for the execution of java scripts and produced outputs which were significantly different from the real world representation of the website when accessed through a browser. Furthermore, it was noted that few modern webpages were loaded completely from simple java scripts that were the only contents in the HTML source. This meant that the data retrieved for these websites was practically useless for purposes of analysis.

To overcome these problems, a headless browser 'HTML-Unit' was used. HTML-Unit can be used to emulate a real browser but not render the output in graphic format on a user-screen. This means that all the background processes of a real browser would be run on a link and thus the entire contents would load and be executed like for a real user. After this occurs, the Discovery engine downloads the entire contents of the webpage HTML source into the database for further analysis and then downloads the entire webpage content (stylesheets, files and images) into a new folder for the webpage inside a predefined directory.

3.4.2 Gathering Metadata

Once this is done, the Discovery engine uses an open DNS whois lookup service called whoapi.com to try to obtain the admin, registrar and registrant information for the domain name of the website in question. It limits the query frequency to administrator specified levels to ensure the tool does not get blocked by the administrators of the service and conforms to the usage policies of the service. This meta-data can be used by investigators to try to classify websites based on owner, registrant country etc. to derive parallel conclusions that are not completely relevant to the functioning of the tool.

The Discovery engine also gathers the data about the technologies used to build the specified website using a service provided by Builtwith.com. This can reveal insights into patterns in technology trends and potential weaknesses that the engine might face in the future.

Both these APIs produce JSON results which are both stored in raw and parsed forms for easier access for the end user.

The webcrawler module of the Discovery engine can be used to crawl flagged websites to a variable depth and download all their contents in a similar fashion, document their content file structure similar to the regular homepages and store them in a separate directory.

Also, a forced periodic re-downloading of webpages can be scheduled and the database will maintain all historic data in the main tables where the webpages and all its subpages are stored.

The tool can be run to seed the initial database from a list of webpages in a simple flat file or individually using the UI developed for the Discovery engine and all the processes can also be scheduled to run at a predefined time which can be randomized as well.

3.4.3 Vocabulary Miner

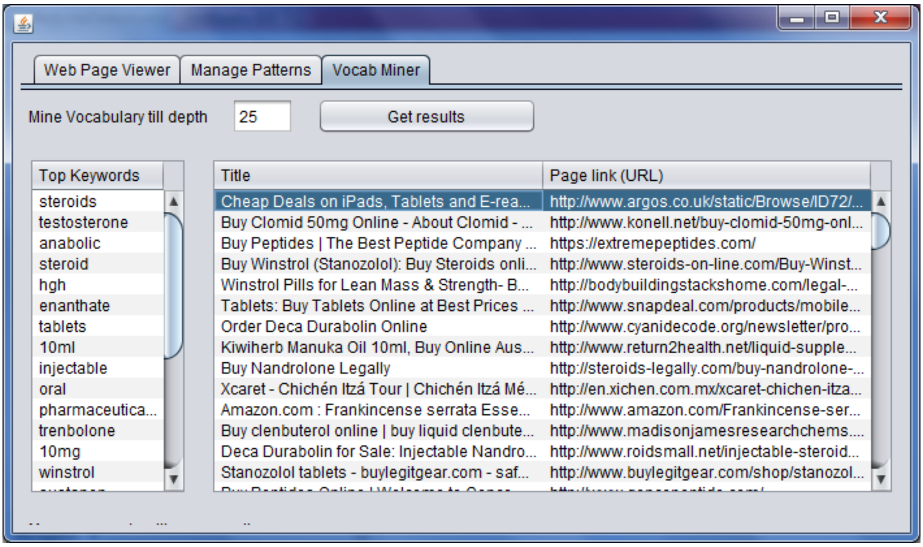
A VocabMiner is used to lookup all the words used in the website database based on the contents rendered from the HTML component. The miner module also counts the number of times each word occurs in the website database across all webpages. This mapping is then ordered and the list of keywords that recur most frequently are singled out. These keywords are then filtered against the blacklisted vocabulary list to remove common English connectives, words used in sentence formation and common words that occur across all websites in general.

A variable number of highest frequency keywords that remain can then be searched over commonly used search providers such as Google. A list of the top results for each of these keywords can then be fed back into the system as newly discovered websites.

The user interface for this is a simple page which provides the user the ability to enter the top ‘x‘ number of keywords to be entered and returns the keywords and their results in a presentable format as shown in the Figure 3.1.

3.5 Results and Evaluation

In this section we discuss the results obtained using the tool described and then attempt to verify the hypothesis in Section 3.3 stated by evaluating them.



Mine Vocabulary till depth 25

Top Keywords
steroids
testosterone
anabolic
steroid
hgh
enanthate
tablets
10ml
injectable
oral
pharmaceutica...
trenbolone
10mg
winstrol

Title	Page link (URL)
Cheap Deals on iPads, Tablets and E-rea...	http://www.argos.co.uk/static/Browse/ID72/...
Buy Clomid 50mg Online - About Clomid - ...	http://www.konell.net/buy-clomid-50mg-onl...
Buy Peptides | The Best Peptide Company ...	https://extremepeptides.com/
Buy Winstrol (Stanozolol): Buy Steroids onli...	http://www.steroids-on-line.com/Buy-Winst...
Winstrol Pills for Lean Mass & Strength- B...	http://bodybuildingstackshome.com/legal-...
Tablets: Buy Tablets Online at Best Prices ...	http://www.snapdeal.com/products/mobile...
Order Deca Durabolin Online	http://www.cyanidecode.org/newsletter/pro...
Kiwiherb Manuka Oil 10ml, Buy Online Aus...	http://www.return2health.net/liquid-supple...
Buy Nandrolone Legally	http://steroids-legally.com/buy-nandrolone-...
Xcaret - Chichén Itzá Tour | Chichén Itzá Mé...	http://en.xichen.com.mx/xcaret-chichen-itza...
Amazon.com : Frankincense serrata Esse...	http://www.amazon.com/Frankincense-ser...
Buy clenbuterol online | buy liquid clenbute...	http://www.madisonjamesresearchchems....
Deca Durabolin for Sale: Injectable Nandro...	http://www.roidsmall.net/injectable-steroid...
Stanozolol tablets - buylegitgear.com - saf...	http://www.buylegitgear.com/shop/stanozol...

Figure 3.1: *Screenshot of the Vocabulary miner tool user interface*

The preliminary investigation is based on the seed dataset provided by DBI which included 423 websites that were verified at the time of creation to contain data that was relevant ie. assured to manually be websites that were selling or appear to be selling prescription-only pharmaceutical products illegally. After removing a few duplicates we arrived at a total list of 412 websites.

After this filtering, it was discovered that only 318 of these websites were actually online at the time of running the tool.

We handle the evaluation of each hypothesis stated under a separate subsection based on the data from these 318 websites.

3.5.1 Hypothesis 1

From the database of 318 available websites that were downloaded into the test database, the vocabulary miner was used to retrieve the keywords of highest frequency.

This was followed up with a manual categorization of the first 300 keywords to remove commonly occurring English language connectives and keywords related to common online shopping terms. A total of 83 words were removed as common

English connectives, prepositions etc. and a total of 126 words were removed as specific irrelevant language for our business goal of pharmaceutical products. The fact that this number includes many words in both their singular and plural form individually is noteworthy. Without them, the total number of blacklisted words would come down even further. For a complete list of these keywords, refer to the appendix at A.2.

Once the blacklisted keywords were applied, the results returned included high quality keywords but search results were still not on par with expectations as they appeared generic and misguided in most cases.

When the top keywords from the vocabulary miner were then combined with the custom keywords of ‘buy‘ and ‘online‘ the results seemed much better than initially expected.

An attempt to retrieve the top search results for the top 10 keywords returned a total of 79 webpage results.

The page headings and their link structures were analyzed to check if the content was interesting from the case at hand as manually examining these pages would be a considerable amount of effort invested. Analyzing these manually revealed that the 66 of the total of 79 webpages returned contained page headings that revealed potentially interesting results. This is roughly 84% of the results returned. It also shows that if they were fed into the identification engine would give great results with a very high likelihood.

An attempt to retrieve the top search results for the top 25 keywords returned a total of 218 webpage results.

Based on a similar analysis of the headings and the web pages url structure, we could see that 186 of the total 218 pages returned contained page headings that revealed potentially interesting results. This shows that roughly 85% of the results discovered in this method were useful. We can thus conclude that the increase from 10 to 25 top keywords retained the quality of the results returned.

The study can be continued to find out the maximum number of search results that need to be mined and searched with before the quality starts to deteriorate but that would need much more manual evaluation than resources permit in this case. But it is proven that the approach works with reasonably good results and can be taken up for further research on its own to study the evolution of keywords over a period of time. Also the effect of using other custom keywords in addition to the ones used in the study above can provide interesting results on their own.

The entire list of keywords that were returned and the weblinks along with their classification can be found at A.1 in the appendix.

Hence our hypothesis of using a standard search engine with a filtered set of high frequency words yielding websites that we are interested in is proved true if we consider using these keywords in combination with user defined custom keywords.

3.5.2 Hypothesis 2

The tool was used to automatically download all pages from the seed list into the database and then an attempt to crawl them to a depth of 1 was made. Conservative estimates showed that each of these webpages on an average contained 73 outgoing links in them to crawl to. It is clear that over 50% of these webpages had more than 52 outgoing links and less than 10% of them had less than 10 outgoing links. The graph in Figure 3.2 shows the data that this assumption based on.

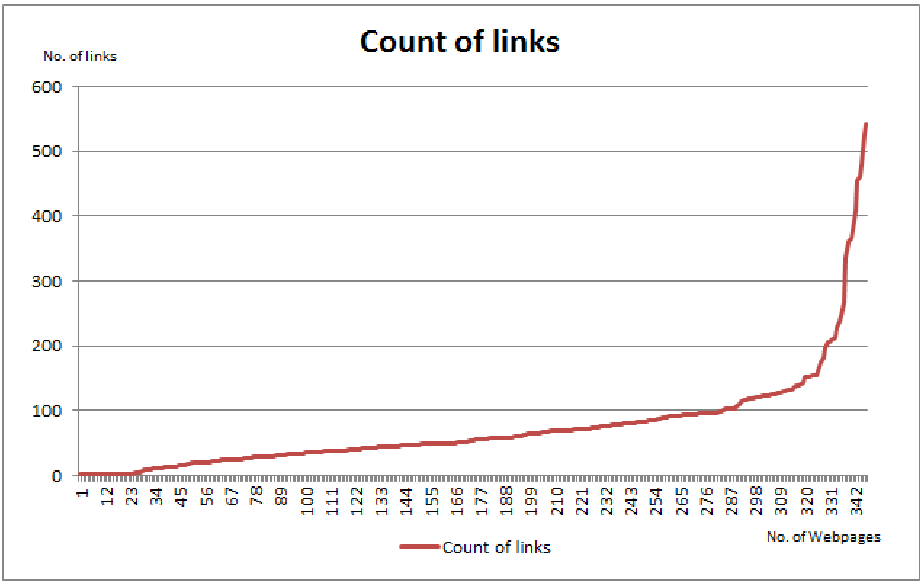


Figure 3.2: *Graph of no. of links in each webpage to no. of webpages with those many links*

Based on manual observations, it is noted that the content structure in many of the subpages that originate from the homepage are similar. The title bar, recommended product list, suggested products and the navigation pane remain constant with only the main content section of the page changing in most subpages.

Hence, if the depth was increased merely to 2 from 1, it is considered reasonable to assume that these numbers hold for all further subpages downloaded too. Through elimination of commonly occurring white-listed pages and removing duplicates that have already been cataloged, it would be possible to reduce these numbers but it would still consume considerably amount of resource nevertheless. Estimating the time required to gather all the data required for a single webpage to be downloaded as roughly 2 minutes, this would clearly be an explosion of effort into downloading webpages that clearly might not need any attention from us.

Besides after analyzing a random sample of about 40 different websites, it was clearly observable that most of these webpages had indicative content on their front pages as is expected from our manual observations on the downloaded webpages. It also seemed reasonably intuitive that the front page of a webpage contains data that depicts the content of the entire website to a large extent.

From observations in case of websites with reasonable number of links in them (close only to 50% of the average of 73 links), it is clearly visible that these are invariably links to individual products webpages which would add significant value to categorization and identification efforts from any automated tool. Thus it offers good reason to invest resources into downloading cataloging and analyzing the information from these specific subpages.

It was also a recurring theme that they lead to other ‘certifying‘ websites that could vouch for their credibility. These websites could potentially provide sources to others of interest that are yet to be discovered.

From our testing, the number of instances when an unrelated website lead to a website that was of interest was close to zero.

Hence our hypothesis of crawling only webpages that were related reducing the workload of the system significantly without much loss to the actual quality of information gathered is proved with good certainty

Chapter 4

Identification Engine

4.1 Introduction

This chapter deals with the detailed description of the features classified under the Identification Engine part of the tool, its purpose and how the tool handles the feature in its implementation.

A clear hypothesis is stated based on expectations of how the tool is built and what output is expected to the effect of the said hypothesis.

We then proceed to document the results and arrive at a conclusion of whether the hypothesis was proved to be true or false.

4.2 Background and Purpose

This part of the project revolves around the ability of the tool to derive measurements from the data that is gathered by the Discovery engine and user feedback to help make decisions on the ranking a page receives when it is added to the database for consideration.

A higher ranked web page should directly correlate with the probability of the webpage containing content that we are interested in based on the seed data set and all further datasets that receive a good user vote. Every page presented to the user can receive either an "upvote" or a "downvote". Upvoted websites are the ones that we are interested in matching from further results and the downvoted websites are the ones we are not.

Due to resource limitations, an investigator might only have time to review manually, a total of 20 websites per day. The database could contain a potential list of many hundred websites that it obtains from various sources, many of which might not actually contain content that the investigator finds interesting. Thus it might take many weeks to run through all the websites in his database before he gets a few hits and end up with an even bigger database by then!

It is therefore important for the investigator to get his eyes on websites that are highly likely to have interesting content first, followed by the other entries in the database.

There are multiple considerations in this section of the tool that make it especially tricky for the strategies that we can uptake to get this tool working.

The first of which is the automation aspect. We would need the tool to require minimal user intervention to retain its accuracy in identifying relevant websites. For this, a strategy similar to the one adopted in the discovery engine is taken up but with a simple change. Instead of building a universal dictionary from the websites database and then comparing that against the new entries, we go with comparing each new website entry to be flagged against all other already up-voted websites in our database. This also aids the Classification Engine that we will discuss later in another chapter of the report.

The second consideration is to be able to quantify the similarity between pages using some meaningful metric. It must not only be able to express this accurately but also for a wide array of parameters that data will be collected for by the Discovery Engine. As a means to achieve this, it becomes important to identify the most appropriate metric to measure the "distance" between two given websites. It must be precise, expressive and also proven and accepted mathematically.

Hence if we use a number of these parameters to derive the datasets and compute the distance between all pairs of websites in the database for each parameter, we will be able to measure how far each of these websites are to each other in terms of that particular parameter being considered. This can then be used to study which parameters yield meaningful results and be used to further obtain better identification capabilities in the tool.

The third consideration is that of user defined patterns to enable slicing data results into more business focused results. For instance, an investigator can define patterns to match specific drugs and pharmaceutical companies. They could then lookup the top results obtained by the methods described above but only those results that match this specific pattern or in other words, "top results" for that specific drug or Pharmaceutical Company. The idea provides an extremely flexible indexing mechanism for results in different ways and interpolating the results from automatic ranking with business needs.

4.3 Hypothesis

The following parameters for datasets, if used in the Jaccard distance method described earlier, could yield potentially valuable identification results that can be used to rank website based on their content.

- File names and folder structure.
- Builtwith information of the website (technologies used to build the website obtained from the service by builtwith.com).

- Vocabulary of the author of a website.
- HTML tag count and structure.
- Phrases and sentence matches in the web-page.

These parameters are based entirely on observations from how a human investigator performs these activities in real life either consciously or subconsciously.

4.4 The Tool

The primary task at hand in this section of the tool is to extract relevant and processable information from the raw data that the Discovery engine has created and use an appropriate parameters to rank the pages based on their relevance.

The capabilities of the identification engine can be broadly classified into the following.

- Calculating the distance value for various parameters of the websites based on the metric defined.
- Provide a ranking scheme.
- Provide support for user defined patterns described earlier.

These are elaborated further in the subsections below.

4.4.1 The Metric for Ranking

One of the primary tasks of the Identification engine, is to rank webpages based on their similarity to other pages in the database. In order to quantify their similarity to other webpages, a metric for measuring "distance" between webpages to express similarities is required.

During the high level study of literature study we came across a research article, that classified fake escrow and financial scam websites on the grounds of structural and content similarity. Use of the metric and the parameters they had chosen was demonstrated to be of high value[8]. Since this study was highly relevant and similar to our discussion, it was identified that the Jaccard distance which was used as a similarity metric in that research as a potential candidate.

The Jaccard distance between two sets S and T is defined as J(S, T), Where J(S,T) = 1 − (| $S \cap T$ | / | $S \cup T$ |)

To elaborate, if S is the set of words used in a website w1.com and T is the set of all words used in website w2.com, then the numerator is the number of words common between the two websites w1 and w2 while the denominator is the total number of words in the vocabulary of both websites put together. Thus we can represent the distance between two sets as value ranging from 0 to 1 when 0 indicates that both sets are completely identical while 1 indicates they are completely distinct. When we use this metric with various measures from the Discovery engine we arrive at a number of interesting results [20].

Further research on metrics that measure similarity(specially in text based sets as in plagiarism detection) revealed a few other options which were analysed.

The Cosine coefficient takes into account not only the actual objects in a set but also their frequencies in computing similarity between the sets. This produces more accurate results than the Jaccard distance but is significantly more expensive to compute. The performance gains in terms of detection of actual plagiarised text also only slightly improved for a steeper increase in computational demand during research[11].

It was also seen that the Jaccard distance was commonly used in plagiarism detection software as a pre-filter in many cases owing to its high accuracy to performance ratio[21]. This was a very key point for this study since the database of websites that we would need to calculate similarity for was expected to be large.

The Dice coefficient is very similar to the Jaccard distance and hence performs equally faster than the other metrics discussed[22]. However this metric was not taken up because it does not satisfy the triangle inequality property[]. This would mean that the distances between two unrelated set of points could not be compared to infer any meaning about the points within the set that were directly compared.

Thus the Jaccard distance was finalized as the metric to be used throughout the study. But a cosine coefficient based metric is also recommended with the Jaccard distance as a pre-filter to improve performance as part of future work to catch more sophisticated forgery.

A modified version of the Jaccard Distance is used by adding the frequency of data elements to the actual elements of the set in certain cases during the tool. This retains the same formula of computation of distance but takes into account the frequency of the elements used in the compared sets as well. This is different from a Cosine coefficient though as in the later, similar but not same frequencies would also be detected!

When a website is added to the database, the tool computes the Jaccard distance between all websites and the newly created website for a number of different parameters defined in the Hypothesis section.

Though the file structure, builtwith information and HTML tag counts are simple enough to understand, the vocabulary based distance computation is handled a little differently. The vocabulary is calculated as the unique set of words that are used in each website. This is achieved by picking the rendered text on every HTML component in the webpage. The rendered text is obtained by using an HTML parser, JSoup to parse the HTML and obtain the text. This text is then broken down into words by using white space as a delimiter after removing common special characters such as commas, fullstops, exclamation marks and question marks. Then a filtering of common words such as connective and prepositions occurs based on entries in a blacklisted word table in the schema.

Once this is done, the Jaccard distance between these two sets is computed and stored in the database for all pairs of websites in the database that are available at the time of processing.

The Jaccard distance for HTML tags between websites is computed based on the sets defined by data entries that are a simple combination of HTML tag name followed by an integer count of the tags occurrence in the webpage as in previous research for easier reproducibility [8]. This is also obtained using the JSoup HTML parser. File name and structure based Jaccard distance is computed on the result from traversing the folder structure of the downloaded webpage content created by the Discovery engine.

As for phrases used in the webpages, the text obtained from the JSoup parser as render text for each HTML component is used directly without breaking them down into words like for the vocabulary based comparison described earlier.

For the purpose of using builtwith data as a parameter for Jaccard distance computation, the technology name for each entry from builtwith for each webpage is used for the computation.

Using inputs from the vocabulary based Jaccard distance, a ranking for the webpages is created for the website as part of two categories. The first being the global rank, or the rank of the webpage out of all the user upvoted webpages.

The second is the ranking of the webpage out of all the webpages that are yet to be voted on. This can be used to look into new websites that require voting in the appropriate order by an investigator.

4.4.2 Ranking Scheme

The rank for a webpage is calculated based on its similarity to the other webpages in the database in terms of vocabulary. This helps bubble up websites that are very similar to each other and weed out ones that are very distinct and stand out from the rest of the database. A filter on only considering webpages that are upvoted can cause this ranking to be of much more value. An alternate rank based on similarity to a global dictionary of upvoted websites can also be used as a means to reach the same goal.

To elaborate, the rank for a website is calculated as the position of the webpage in a sorted list of minimum Jaccard distance for its vocabulary compared against the vocabulary of each other website in the database. The second method to determine rank would be to calculate individual Jaccard distances for each website against the global upvoted dictionary and sort this list.

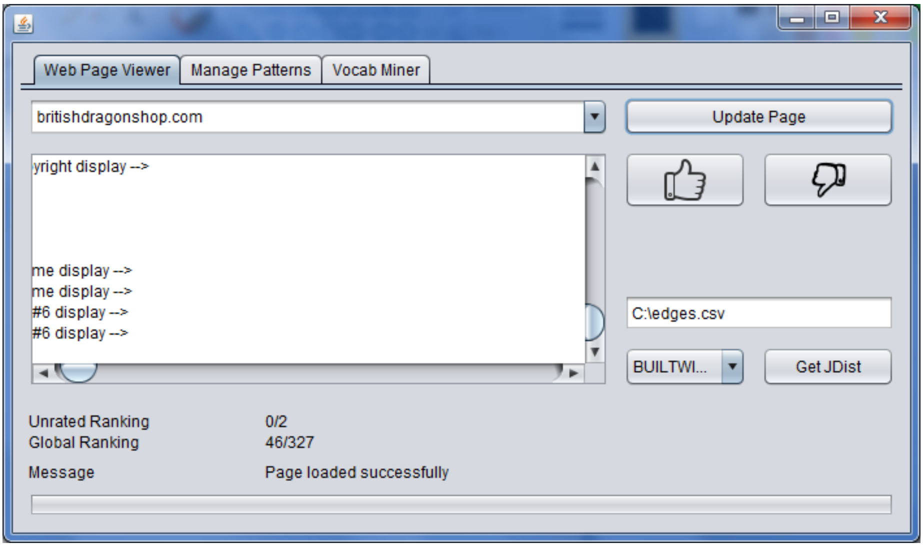
The key difference between the two methods of determining rank are that the former method would rank a page higher if it resembled even one other page very much. While the later method would probably rank a page that is closer to a number of pages in the database or in other words similar to the average of the entire database.

Both serve a unique purpose and it was chosen to implement only the former since it can be reused in the Classification Engine (Chapter 5) and time limitation. But it is highly recommended that the other approach be taken up in future work as it appears to promise very insightful analysis.

This information and the actual content of the webpage together can be used to upvote or downvote a particular website using the UI. The simple user interface for this purpose in the tool can be seen in Figure 4.1.

4.4.3 User Defined Patterns

User defined patterns are defined and managed through the use of pattern groups. Pattern groups are a collection of individual patterns which are simply keywords that are directly matched in the source code. Individual patterns are



britishdragonshop.com

```
yright display -->



me display -->
me display -->
#6 display -->
#6 display -->
```

C:\edges.csv

BUILTWI...

Unrated Ranking 0/2
Global Ranking 46/327

Message Page loaded successfully

Figure 4.1: *Screenshot of the Overview page of the tool*

matched in the source of a webpage directly and the results are indexed in the database. The key use of this feature of the tool is to provide the ability to classify webpages based on specific user defined criteria. Patterns such as ‘sell‘ or ‘buy‘ can correspond to a pattern group ‘Sale‘. Similarly, patterns such as ‘Novo nordisk‘ and ‘Astrazeneca‘ can correspond to a pattern group called ‘Companies of Interest’.

Once these pattern groups and their corresponding patterns are defined, it is possible for us to run a matching algorithm against our database and extrapolate results for queries such as, what are the top 10 websites that are replicated and active the most ‘Selling‘ products from ‘Companies of Interest‘.

The Figure 4.2 shows the simple user interface from the tool for this particular feature.

4.5 Results and Evaluation

In this section we discuss the results obtained using the tool described and then attempt to verify the hypothesis stated by evaluating them.

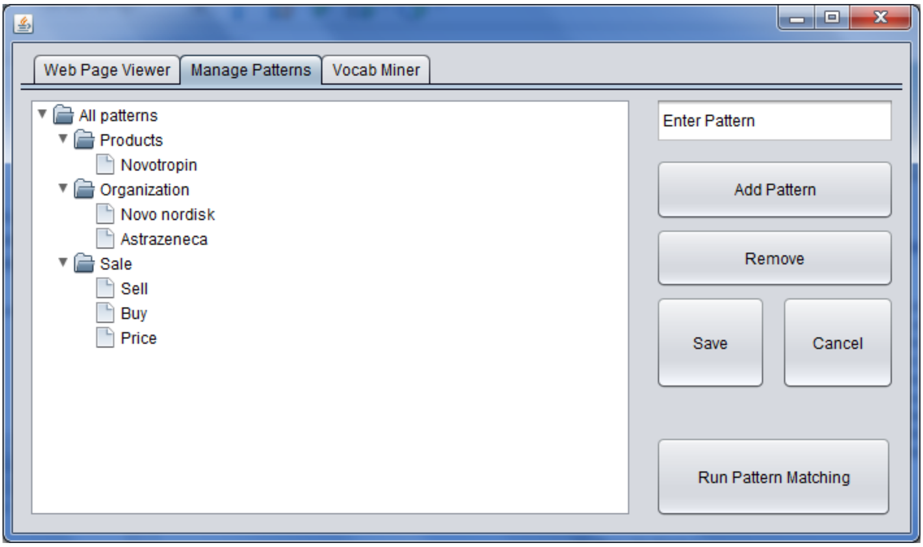


Figure 4.2: *Screenshot of the User defined pattern section of the tool*

The preliminary investigation is based on the seed dataset provided by DBI which included 423 websites that were verified at the time of creation to contain data that was relevant, that is assured to manually to be websites that were selling or appear to be selling prescription-only pharmaceutical products. After removing a few duplicates we arrived at a total of 412 websites in the list.

After this filtering, it was discovered that only 318 of these websites were actually online at the time of running the tool.

Since the identification potential of the parameters of file structure, HTML tag count and phrases and sentences as defined in the hypothesis section 4.3 have already been discussed thoroughly as part of previous research, it was decided that the work will focus primarily on the vocabulary information while the classification potential of the builtwith information will be dealt with in the appropriate chapter [8].

All the websites that were originally part of the list provided by DBI were upvoted to provide the grounds for evaluation of further websites by the engine.

In the further subsections of the results and analysis, we classify the tools ranking capabilities in the context of three types of webpages.

Firstly the analysis of the existing seed database of webpages from DBI. Secondly the analysis of ranking provided by the tool for websites that do not conform to the existing seed database and if they are ranked low. Finally the analysis of ranking provided by the tool for websites that conform to the existing seed database and if they are ranked high.

4.5.1 Analysis of Existing Database

In this section we discuss the analysis of the ranking of the webpages in the system and what they revealed about the webpages in general.

A distribution graph of the Jaccard distance by ranking shows us that four distinct classes of websites as can be seen in the Figure 4.3.

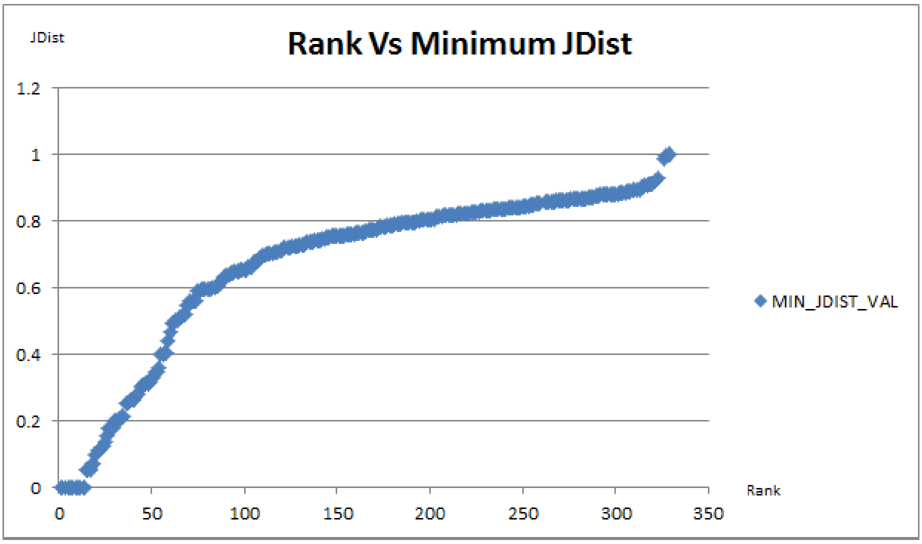


Figure 4.3: *A graph plot of the Ranking of webpages against their lowest Jaccard distance from all pairs of websites for their vocabulary.*

The lowest region or the first 15 ranks which have a Jaccard distance of 0 are basically webpages that had an exactly identical entry in the database. Further analysis revealed that this was due to erroneous entries in the database from the duplicate user input that were because of a few letters being lower and upper case. They can be discarded from our analysis.

The next region comprises the steep curve that indicates a sharp increase in the Jaccard distance over a small set of ranks. Upon manual analysis these consisted of very similar websites that had good amount of duplicate content in them. These were websites that were highly likely to have been copied from one another and had a good probability for having been created by the same author.

The following region comprises the stable Jaccard distance of up to 0.85 where most websites lay and the curve stabilizes. These were webpages which didn't have too many very similar webpages but still had a reasonable amount of data common between each other. This is hypothesized to be because of the small size of the dataset initially taken into consideration. If the number of entries in the database are increased we would probably segregate these websites much better than now.

The last section of the graph consists of webpages with extremely high minimum Jaccard distances. These were predominantly discovered to be pages that were erroneously categorized as available when they actually were either captcha pages, or redirect pages, or page not found default pages. Hence they can also be ignored.

Looking at these figures it appears fair to assume that the categorization would work much better with a ranking based on the global dictionary built from only the upvoted websites. It is also worth mentioning that the proper validation and cleaning of data before being entered into the database is of utmost importance to maintain realistic figures in this section.

4.5.2 Analysis of Non-conformant Entries

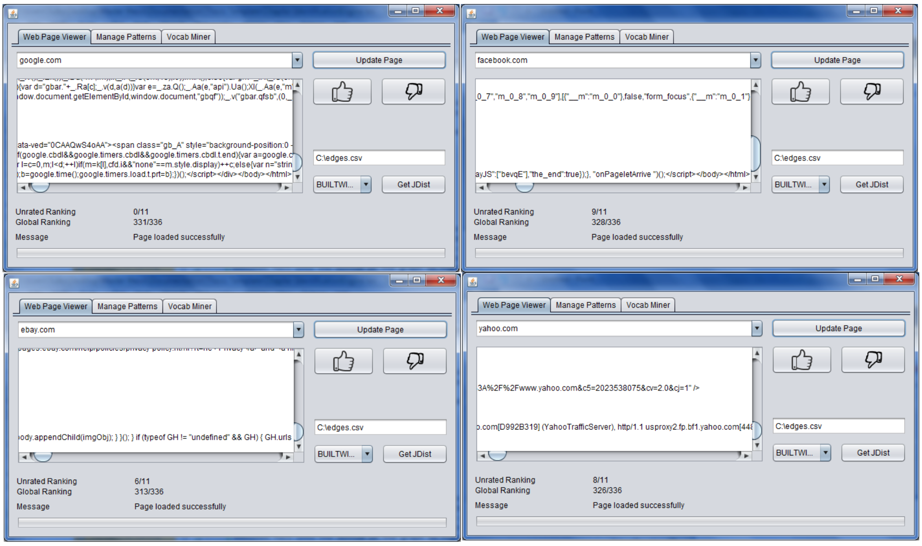
In this section we discuss the results that were obtained for entering websites that were expected to be ranked lower because they were manually verified to contain irrelevant information from our existing database.

An attempt was made to add the following pages that were not related to what we were looking for and their global ranking was logged.

- google.com - 325/328
- facebook.com - 322/329
- amazon.com - 321/330
- ebay.com - 310/331
- yahoo.com - 323/333

A list of the top websites from Alexa ranking were chosen for this and surprisingly, even the E-Retail websites such as amazon and ebay ranked low in the bottom 5% of the ranking despite commonalities in the fact that they sell products to their customers similar to many f the websites that we were interested in. So did a webpage like yahoo.com, with its large homepage content which might have thrown the ranker off course. This demonstrates how effective the method is in segregating websites irrelevant to our discussion very effectively. We can clearly see that all of these websites fall into the bottom of the graph in Figure 4.3 as expected.

Figure 4.4 shows the results from the tool for these websites.



google.com

```
{var d="gbar."+_.Ra[c];_.v(d,a(d))}var e=_.za.Q();_.Aa(e,"api").Ua();XI(_.Aa(e,"m
dow.document.getElementById,window.document,"gbqf"));_.v("gbar.qfsb",(0,_

ata-ved="0CAAQwS4oAA"><span class="gb_A" style="background-position:0 -
f(google.cbdl&&google.timers.cbdl&&google.timers.cbdl.t.end){var a=google.c
r l=c=0,m;l<d;++l)if(m=k[l],cfd.i&&"none"==m.style.display)++c;else{var n="strin
;b=google.time();google.timers.load.t.prt=b};})();</script></div></body></html>
```

C:\edges.csv

Unrated Ranking 0/11
Global Ranking 331/336
Message Page loaded successfully

facebook.com

```
0_7","m_0_8","m_0_9"],[{"__m":"m_0_0"},false,"form_focus",{"__m":"m_0_1"}

ayJS":["bevqE"],"the_end":true});}, "onPageletArrive ")();</script></body></html>
```

C:\edges.csv

Unrated Ranking 9/11
Global Ranking 328/336
Message Page loaded successfully

ebay.com

```
ody.appendChild(imgObj); } }(); } if (typeof GH != "undefined" && GH) { GH.urls
```

C:\edges.csv

Unrated Ranking 6/11
Global Ranking 313/336
Message Page loaded successfully

yahoo.com

```
3A%2F%2Fwww.yahoo.com&c5=2023538075&cv=2.0&cj=1" />

.com[D992B319] (YahooTrafficServer), http/1.1 usproxy2.fp.bf1.yahoo.com[44
```

C:\edges.csv

Unrated Ranking 8/11
Global Ranking 326/336
Message Page loaded successfully

Figure 4.4: *Screenshot of tool with the non-conformant websites entered.*

It also proves the strong coupling between the vocabulary of the specific websites that we are interested in the database.

4.5.3 Analysis of Conformant Entries

In this section we discuss the results that were obtained for entering websites that were expected to be ranked higher because they were manually verified to contain interesting information.

An attempt was made to find new websites that were not part of the list provided by DBI and were added to the database with their ranks logged.

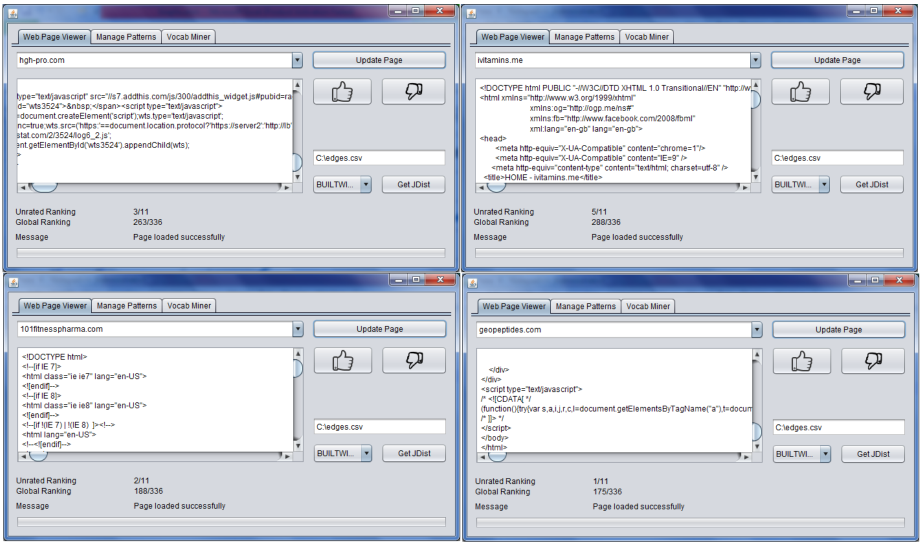
- hgh-pro.com - 260/333
- shop.hgh-pro.com - 261/333
- ivitamins.me - 286/334
- 101fitnesspharma.com - 185/335
- geopeptides.com - 175/336

The list of websites to be considered were sourced from the eroids.com forum which provides an excellent source of reviewed webpages that are known to sell counterfeit pharmaceutical products online and from a general Google search.

It is clear that even the worst of these websites are ranked significantly higher than the best non-conformant websites. Some fall under the third region in the graph in Figure 4.3 but few fall right into the second region of the graph.

As a consequence, the unvoted rank for all the conformant webpages are higher than those of the non-conformant webpages which is as expected.

Figure 4.5 shows the results from the tool for these websites.



hgh-pro.com

```
type="text/javascript" src="//s7.addthis.com/js/300/addthis_widget.js#pubid=ra
d="wts3524">&nbsp;</span><script type="text/javascript">
=document.createElement('script');wts.type='text/javascript';
nc=true;wts.src=('https:'==document.location.protocol?'https://server2':'http://lb'
stat.com/2/3524/log6_2.js';
ent.getElementById('wts3524').appendChild(wts);
>
```

C:\edges.csv

Unrated Ranking 3/11
Global Ranking 263/336
Message Page loaded successfully

ivitamins.me

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://w
<html xmlns="http://www.w3.org/1999/xhtml"
          xmlns:og="http://ogp.me/ns#"
          xmlns:fb="http://www.facebook.com/2008/fbml"
          xml:lang="en-gb" lang="en-gb">
<head>
   <meta http-equiv="X-UA-Compatible" content="chrome=1"/>
   <meta http-equiv="X-UA-Compatible" content="IE=9" />
  <meta http-equiv="content-type" content="text/html; charset=utf-8" />
 <title>HOME - ivitamins.me</title>
```

C:\edges.csv

Unrated Ranking 5/11
Global Ranking 288/336
Message Page loaded successfully

101fitnesspharma.com

```
<!DOCTYPE html>
<!--[if IE 7]>
<html class="ie ie7" lang="en-US">
<![endif]-->
<!--[if IE 8]>
<html class="ie ie8" lang="en-US">
<![endif]-->
<!--[if !(IE 7) | !(IE 8)  ]><!-->
<html lang="en-US">
<!--<![endif]-->
```

C:\edges.csv

Unrated Ranking 2/11
Global Ranking 188/336
Message Page loaded successfully

geopeptides.com

```
  </div>
</div>
<script type="text/javascript">
/* <![CDATA[ */
(function(){try{var s,a,i,j,r,c,l=document.getElementsByTagName("a"),t=docum
/* ]]> */
</script>
</body>
</html>
```

C:\edges.csv

Unrated Ranking 1/11
Global Ranking 175/336
Message Page loaded successfully

Figure 4.5: *Screenshot of tool with the conformant websites entered.*

Hence the hypothesis that the Vocabulary of the webpage can be used to identify and rank webpages successfully given a seed list of webpages is proved to be correct and accurate.

4.5.4 Other Thoughts

As described earlier the possibility of ranking websites based on Jaccard distance from a consolidated global dictionary for upvoted websites could potentially yield better results in ranking the webpages that are relevant but it also depends on what is considered important.

In this study, the consideration for a webpage being ranked higher was to be the source of multiple spin off or duplicate websites which could be easily augmented in a single case as described earlier. For this purpose, this ranking provides quite an accurate result but the alternate approach is definitely worth consideration for future work.

Chapter 5

Classification Engine

5.1 Introduction

This chapter deals with the detailed description of the features classified as the Classification Engine as part of the tool, its purpose and how the tool handles the feature in its implementation. This can also be viewed as an advanced feature of the Identification Engine.

A clear hypothesis is stated based on expectations of how the tool is built and what output is expected to the effect of the said hypothesis

We then proceed to document the results and arrive at a conclusion of weather the hypothesis was proved to be true or false.

5.2 Background and Purpose

This part of the project that is more of a subclass of the Identification Engine revolves around the ability of the tool to classify or identify similarities between flagged websites. This directly correlates to real world interests from investigators.

The reason for this components relevance can be explained through the use of a common example from the investigative realm. There are hundreds if not thousands of websites in the open web that market and sell prescription-only drugs to users. Even after identifying these websites, there is an enormous amount of effort that goes into collecting the evidence required to shut these websites down and enforce the decisions from courts especially because of the highly international nature of law enforcement in any case pertaining to crimes over the Internet [14].

This coupled with limited resources dedicated to the discovery, enforcement and maintenance of law in this sector have imposed serious restrains in the way law enforcement acts in such cases.

Hence it becomes vital for law enforcement to look not just at individual websites and see if they are dealing in prescription-only pharmaceutical products but also sister websites that can be processed together and shut down as a result of a single investigative court case rather than individually process each website to make the process less cumbersome and more efficient.

There is certain evidence to believe that criminals dealing in websites that scam or sell illegal articles over the Internet have a strong incentive in creating ‘sister websites‘ that have the similar content but appear different to the user[8]. Some of the key reasons criminals are known to do this for are,

- The business can run even if one of the websites is shut down; simply by redirecting requests to another.
- Scam websites can easily distinguish themselves from other websites which have outlived their effective lifespan.
- Reusing content from the original website is considerably cheaper than creating fresh content for every website that is created.
- Attracting regional audience better through superficial customizations but retaining the core content.

Another benefit of working on classifying websites is to identify the flow in the supply chain of products since it can also possibly bring out supplier-consumer relationships amongst websites and shed more light on further steps to identifying the root of the problem instead of treating mere symptoms that spawn as individual websites.

It might be possible to segregate some sites as retailers of a product while others might be wholesalers that supply many other websites. This can be identified by creating a sort of timelapse of when and where keywords, images and fresh stock start to appear and where they disappear in the database of websites. Of all the websites in our database, the one that posts key new products before any other website does indicates that they acquired these products before the rest. Then the order of the products appearing in the website can help reveal details into where these products flow and how.

Combined with metadata such as hosting information, and real world test purchases that are common in the criminal investigation world today, an incredible amount of deduction is possible using capabilities such as these.

Subsequently, cases pertaining to large wholesalers can be prioritized and processed faster to rapidly control key problem areas in a more agile manner with larger impact on crime.

5.3 Hypothesis

Using the Jaccard distance as a metric it is possible to derive meaningful information about classifying websites based on the following parameters to measure the Jaccard distance based on.

1. BuiltWith information (Technologies used to build the website obtained from the service by builtwith.com).
2. Vocabulary information of the website.
3. Phrases and sentence structures in a website.
4. Folder structure and file names
5. HTML tags and tag counts

5.4 The Tool

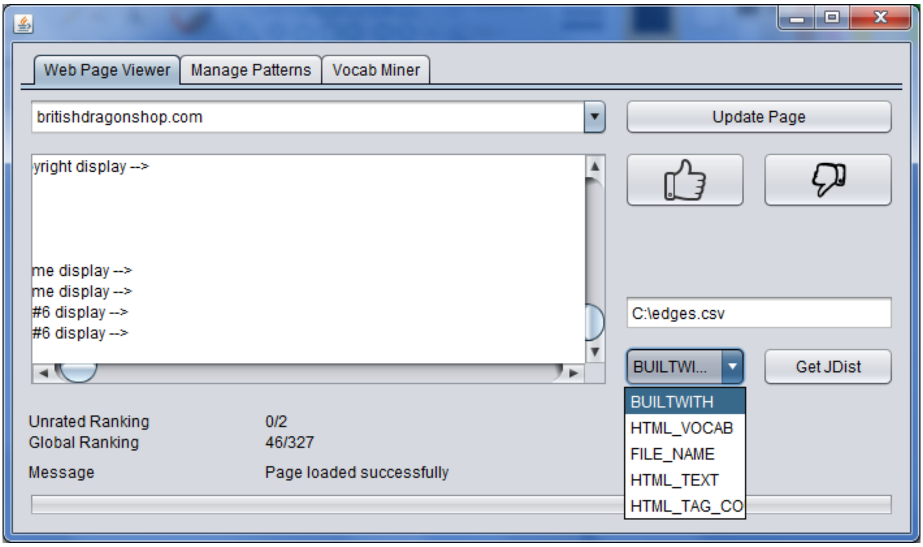
The Classification engines core consists of the utilizing the Jaccard distance computed for each parameter for all pairs of websites in an interesting manner and deriving meaningful subsets of data from the main dataset.

Once a particular website is loaded in the UI, the tool provides a simple way to select the parameter that we wish to extract the Jaccard distance value between all pairs of websites for. This data can be extracted to a csv file to a desired location.

Visualization can be a powerful tool to help a human user understand the vast amounts of data derived from the data mining processes. It helps present data in an attractive, understandable manner and lets users play around and derive highly interesting and out of the box conclusions from large volumes of data [16].

The csv file is tailored to work directly with a data visualization software called Gephi that helps produce dynamic graphs from large dataset products from the tool [3]. Gephi then automatically produces a graph that represents all the websites in the database as points. Websites having lesser Jaccard Distance (similar with respect to the measurement used) are grouped closer to each other while the ones that are not so similar are placed further apart from the one being considered. Gephi can then be used to easily color code websites that fall within a critical Jaccard distance value from the website being considered and make them appear a darker color than the ones that we are not interested in. It can also quickly filter websites that are not relevant because of their high Jaccard distance and present the a visual representation of the data in our DB.

The parameters for which the Jaccard distance can be extracted in the tool right now are the ones that are created when the Identification and Discovery engine first bring a website into the database. If more parameters are added in the future, then they will automatically appear as available parameters to extract and visualize in the tool. This can be seen in tool screenshot in Figure 5.1.



britishdragonshop.com

```
yright display -->




me display -->
me display -->
#6 display -->
#6 display -->
```

C:\edges.csv

Unrated Ranking 0/2
Global Ranking 46/327
Message Page loaded successfully

Figure 5.1: *Screenshot of the tool showing the different extracts available for visualization using Gephi*

The tool is used to extract and visualize the information for all the parameters mentioned in the Hypothesis section and analyzed manually for a sample set of websites in further sections to validate the results.

5.5 Results and Evaluation

In this section we discuss the results obtained using the tool described and then try to verify the hypothesis stated by evaluating them.

We use this section to understand the power of the tool to classify and spot similarities between the websites in the database. And to this end, we use the vocabulary as a parameter and the builtwith dataset as a parameter to compute Jaccard distances between a given website and all others and then visualize it using Gephi.

A number of websites were chosen at random and tested by extracting the all pairs Jaccard distance for the vocabulary and builtwith dataset and then fed into Gephi. Once in Gephi, an attempt was made to study the results by changing the edge weight filter criteria to various values to see if patterns were revealed. If they were reported to be similar, then a manual verification of the website's similarities in a real world browser were made to ensure that the content was actually similar. It also served as a means to document what differences were seen and what other potential leads this could give rise to.

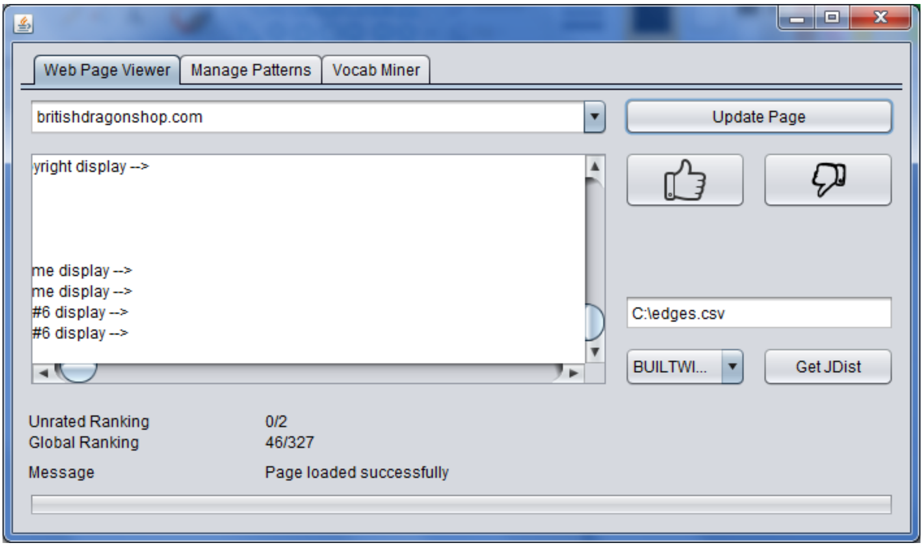
The websites that were tested are classified into three distinct classes based on how likely they were to be replicated in the real world. Replication Likely webpages are those where the likelihood that these websites were actual clones of each other are high. Replication Unlikely webpages are those where the chances of them being clones of other websites in the database are low. Lastly, Partial Replication Likely webpages are those where some but not all content has been replicated off other websites in the database.

5.5.1 Replication Likely Webpages

It is highly likely that the websites that are documented as part of this section are replicas of each other(or have a common root) or have content that is heavily borrowed from one another. We will describe the process involved in obtaining these results and then elaborate what the results actually show.

The website that we consider for this section is one named britishdragonshop.com.

An extract of the HTML_VOCAB is made from the tool for the website britishdragonshop.com as shown in the Figure 5.2. Then it is imported into Gephi as edges and the nodes csv is used for names of the websites in the database.



britishdragonshop.com

```
yright display -->



me display -->
me display -->
#6 display -->
#6 display -->
```

C:\edges.csv

BUILTWI...

Unrated Ranking	0/2
Global Ranking	46/327
Message	Page loaded successfully

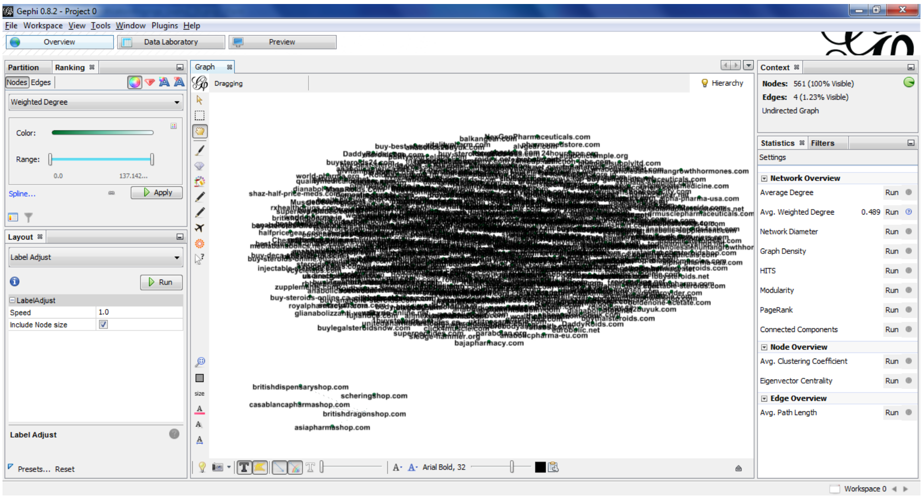
Figure 5.2: *Screenshot of the extraction of builtwith information for BritishDragonShop.com*

From careful observation and experimentation, it is observed that the following steps can be taken up in sequence in Gephi to produce graphs of a certain type. These graphs display information that is very relevant to our study.

All edge weights except the minimum 10% are filtered from the graph to remove vertices from the graph that correspond to high distance values. Thickness of edges is reduced is minimized to reduce clutter. Labels are added to all nodes to show the websites that each node corresponds to and their font size is adjusted to convenience.

Some of the standard layouts offered by Gephi help us transoform the rats nest graphs that we have now into a more meaningful and understandable format. The "Fruchterman Reingold" layout is applied to distribute the nodes evenly followed by "Force Atlas" layout to segregate the set of nodes into the connected bunch and the unconnected bunch. We follow this up with a quick run of the "Label Adjust" layout to show the labels clearly.

A weighted degree based color gradient is applied to the graph to highlight the source node and its connected nodes. At this point, then graph looks as can be found in the Figure 5.3.



Nodes: 561 (100% Visible)
Edges: 4 (1.23% Visible)
Undirected Graph

Avg. Weighted Degree 0.489

LabelAdjust
Speed 1.0
Include Node size ☑

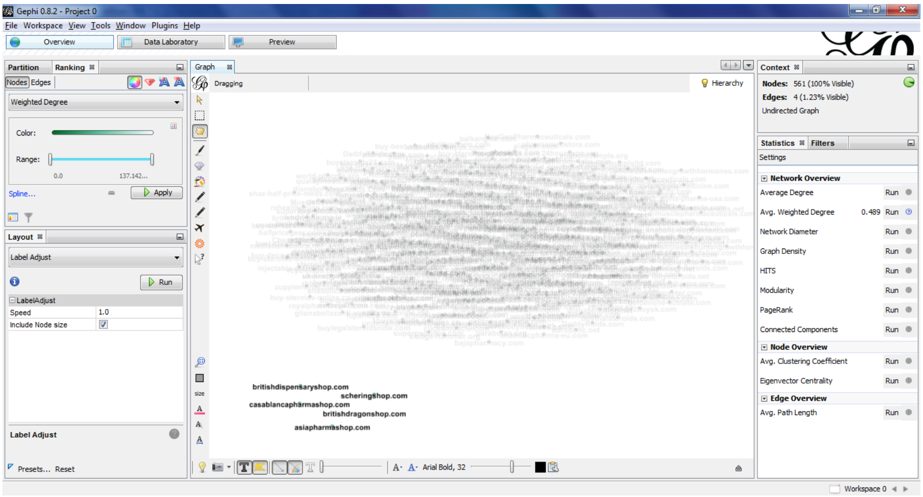
Figure 5.3: *Screenshot of Gephi showing classification of webpages similar to BritishDragonShop.com*

Now if we hover over the source node, which in our case now is britishdragonshop.com, then we can see only those that still have a connection to it through one of the non-filtered edges. This means that they are likely to contain a similar vocabulary of the order of 0.35-0.40 Jaccard distance in this particular case. This can be seen in the Figure 5.4.

It can be seen from these images that the websites that are likely to be similar to britishdragonshop.com are the ones mentioned below.

1. britishdispensaryshop.com
2. scheringshop.com
3. casablancapharmashop.com
4. asiapharmashop.com

To verify this claim manually, a lookup on these websites was made on a Chrome browser to verify the content similarity. It was discovered that there was a remarkable level of similarity between these websites in their principle content as can be seen in the screenshots in Figures 5.8, 5.6, 5.9 and 5.7 below.



Weighted Degree

Range: 0.0 137.142...

Label Adjust

LabelAdjust

Speed	1.0
Include Node size	☑

britishdispensaryshop.com
scheringshop.com
casablancapharmashop.com
britishdragonshop.com
asiapharmashop.com

Nodes: 561 (100% Visible)
Edges: 4 (1.23% Visible)
Undirected Graph

Network Overview

Average Degree		Run
Avg. Weighted Degree	0.489	Run
Network Diameter		Run
Graph Density		Run
HITS		Run
Modularity		Run
PageRank		Run
Connected Components		Run

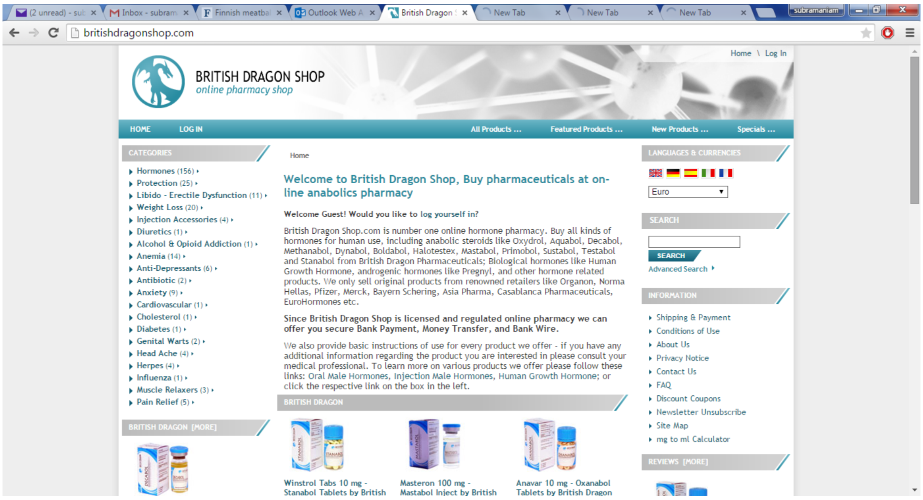
Node Overview

Avg. Clustering Coefficient	Run
Eigenvector Centrality	Run

Edge Overview

Avg. Path Length	Run

Figure 5.4: *Screenshot of Gephi showing classification of webpages similar to BritishDragonShop.com - Source highlighted*



- Hormones (156)
- Protection (25)
- Libido - Erectile Dysfunction (11)
- Weight Loss (20)
- Injection Accessories (4)
- Diuretics (1)
- Alcohol & Opioid Addiction (1)
- Anemia (14)
- Anti-Depressants (6)
- Antibiotic (2)
- Anxiety (9)
- Cardiovascular (1)
- Cholesterol (1)
- Diabetes (1)
- Genital Warts (2)
- Head Ache (4)
- Herpes (4)
- Influenza (1)
- Muscle Relaxers (3)
- Pain Relief (5)

Welcome to British Dragon Shop, Buy pharmaceuticals at on-line anabolics pharmacy

Welcome Guest! Would you like to log yourself in?

British Dragon Shop.com is number one online hormone pharmacy. Buy all kinds of hormones for human use, including anabolic steroids like Oxydrol, Aquabol, Decabol, Methanabol, Dynabol, Boldabol, Halotestex, Mastabol, Primobol, Sustabol, Testabol and Stanabol from British Dragon Pharmaceuticals; Biological hormones like Human Growth Hormone, androgenic hormones like Pregnyl, and other hormone related products. We only sell original products from renowned retailers like Organon, Norma Hellas, Pfizer, Merck, Bayern Schering, Asia Pharma, Casablanca Pharmaceuticals, EuroHormones etc.

Since British Dragon Shop is licensed and regulated online pharmacy we can offer you secure Bank Payment, Money Transfer, and Bank Wire.

We also provide basic instructions of use for every product we offer - if you have any additional information regarding the product you are interested in please consult your medical professional. To learn more on various products we offer please follow these links: Oral Male Hormones, Injection Male Hormones, Human Growth Hormone; or click the respective link on the box in the left.

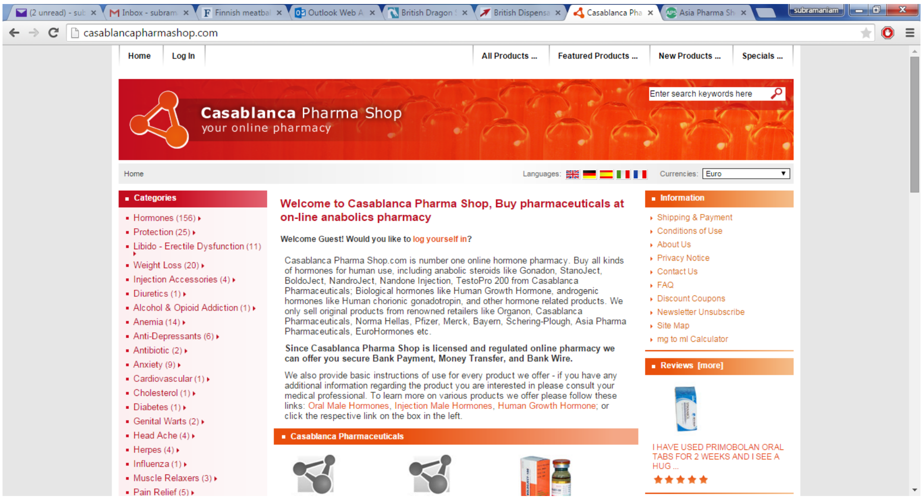
Winstrol Tabs 10 mg - Stanabol Tablets by British

Masteron 100 mg - Mastabol Inject by British

Anavar 10 mg - Oxanabol Tablets by British Dragon

- Shipping & Payment
- Conditions of Use
- About Us
- Privacy Notice
- Contact Us
- FAQ
- Discount Coupons
- Newsletter Unsubscribe
- Site Map
- mg to ml Calculator

Figure 5.5: *Screenshot of BritishDragonShop.com homepage*



Welcome to Casablanca Pharma Shop, Buy pharmaceuticals at on-line anabolics pharmacy

Welcome Guest! Would you like to log yourself in?

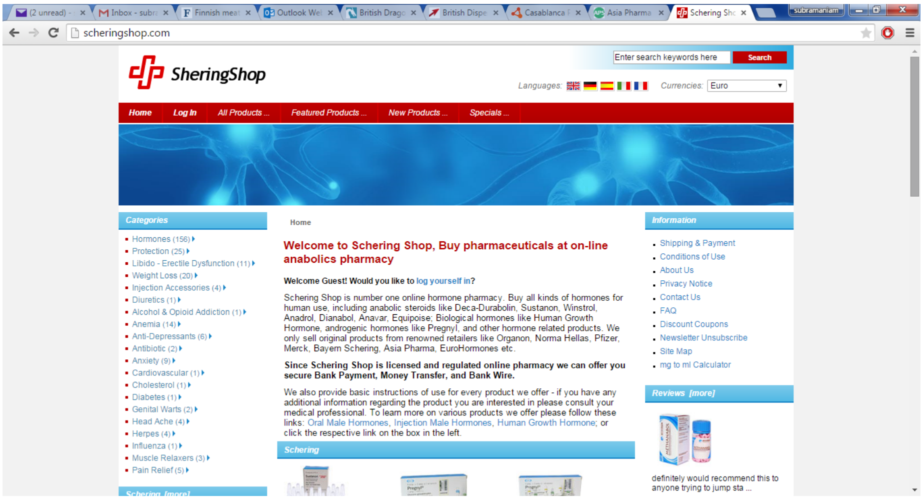
Casablanca Pharma Shop.com is number one online hormone pharmacy. Buy all kinds of hormones for human use, including anabolic steroids like Gonadon, StanoJect, BoldoJect, NandroJect, Nandone Injection, TestoPro 200 from Casablanca Pharmaceuticals; Biological hormones like Human Growth Hormone, androgenic hormones like Human chorionic gonadotropin, and other hormone related products. We only sell original products from renowned retailers like Organon, Casablanca Pharmaceuticals, Norma Hellas, Pfizer, Merck, Bayern, Schering-Plough, Asia Pharma Pharmaceuticals, EuroHormones etc.

Since Casablanca Pharma Shop is licensed and regulated online pharmacy we can offer you secure Bank Payment, Money Transfer, and Bank Wire.

We also provide basic instructions of use for every product we offer - if you have any additional information regarding the product you are interested in please consult your medical professional. To learn more on various products we offer please follow these links: Oral Male Hormones, Injection Male Hormones, Human Growth Hormone; or click the respective link on the box in the left.

I HAVE USED PRIMOBOLAN ORAL TABS FOR 2 WEEKS AND I SEE A HUG ...

Figure 5.6: *Screenshot of CasablancaPharmaShop.com homepage*



Welcome to Schering Shop, Buy pharmaceuticals at on-line anabolics pharmacy

Welcome Guest! Would you like to log yourself in?

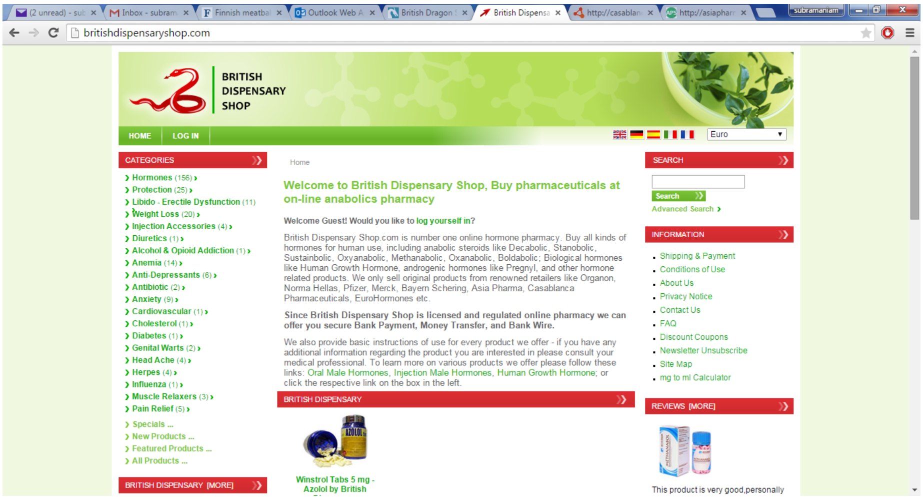
Schering Shop is number one online hormone pharmacy. Buy all kinds of hormones for human use, including anabolic steroids like Deca-Durabolin, Sustanon, Winstrol, Anadrol, Dianabol, Anavar, Equipoise; Biological hormones like Human Growth Hormone, androgenic hormones like Pregnyl, and other hormone related products. We only sell original products from renowned retailers like Organon, Norma Hellas, Pfizer, Merck, Bayern Schering, Asia Pharma, EuroHormones etc.

Since Schering Shop is licensed and regulated online pharmacy we can offer you secure Bank Payment, Money Transfer, and Bank Wire.

We also provide basic instructions of use for every product we offer - if you have any additional information regarding the product you are interested in please consult your medical professional. To learn more on various products we offer please follow these links: Oral Male Hormones, Injection Male Hormones, Human Growth Hormone; or click the respective link on the box in the left.

definitely would recommend this to anyone trying to jump sta ...

Figure 5.7: *Screenshot of ScheringShop.com homepage*



Welcome to British Dispensary Shop, Buy pharmaceuticals at on-line anabolics pharmacy

Welcome Guest! Would you like to log yourself in?

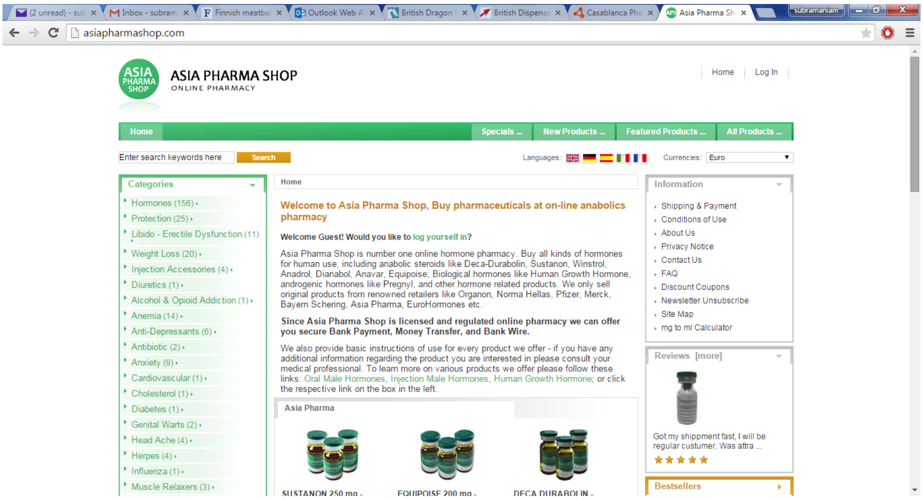
British Dispensary Shop.com is number one online hormone pharmacy. Buy all kinds of hormones for human use, including anabolic steroids like Decabolic, Stanobolic, Sustainbolic, Oxyanabolic, Methanabolic, Oxanabolic, Boldabolic; Biological hormones like Human Growth Hormone, androgenic hormones like Pregnyl, and other hormone related products. We only sell original products from renowned retailers like Organon, Norma Hellas, Pfizer, Merck, Bayern Schering, Asia Pharma, Casablanca Pharmaceuticals, EuroHormones etc.

Since British Dispensary Shop is licensed and regulated online pharmacy we can offer you secure Bank Payment, Money Transfer, and Bank Wire.

We also provide basic instructions of use for every product we offer - if you have any additional information regarding the product you are interested in please consult your medical professional. To learn more on various products we offer please follow these links: Oral Male Hormones, Injection Male Hormones, Human Growth Hormone; or click the respective link on the box in the left.

This product is very good,personally

Figure 5.8: *Screenshot of BritishDispensaryShop.com homepage*



Welcome to Asia Pharma Shop, Buy pharmaceuticals at on-line anabolics pharmacy

Welcome Guest! Would you like to log yourself in?

Asia Pharma Shop is number one online hormone pharmacy. Buy all kinds of hormones for human use, including anabolic steroids like Deca-Durabolin, Sustanon, Winstrol, Anadrol, Dianabol, Anavar, Equipoise; Biological hormones like Human Growth Hormone, androgenic hormones like Pregnyl, and other hormone related products. We only sell original products from renowned retailers like Organon, Norma Hellas, Pfizer, Merck, Bayern Schering, Asia Pharma, EuroHormones etc.

Since Asia Pharma Shop is licensed and regulated online pharmacy we can offer you secure Bank Payment, Money Transfer, and Bank Wire.

We also provide basic instructions of use for every product we offer - if you have any additional information regarding the product you are interested in please consult your medical professional. To learn more on various products we offer please follow these links: Oral Male Hormones, Injection Male Hormones, Human Growth Hormone; or click the respective link on the box in the left.

Got my shippment fast, I will be regular custumer. Was attra ...

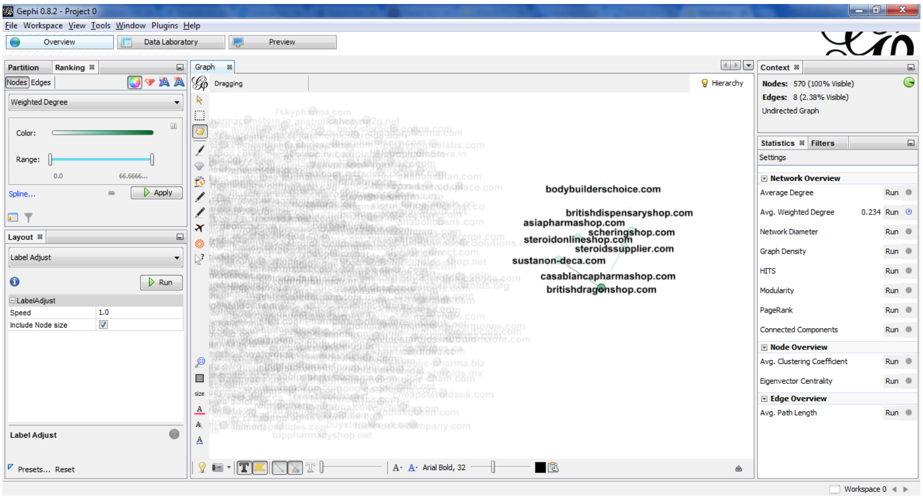
★★★★★

Figure 5.9: *Screenshot of AsiaPharmaShop.com homepage*

It is very clear that these websites only have cosmetic changes in them and that their primary content and content structure bear a striking resemblance. When a deeper dig into the actual numbers was made to identify the source of even the small deviation, it was discovered that the only reason the Jaccard distance was not completely zero was because of the dynamic contents in the homepage of each website. These were the top selling product categories which varied from one website to another and the actual products that contributed to these. Some of these appear random, while others seem to be based on sales trends.

As such it is these that contribute to even the small deviations in Jaccard distance for the vocabulary of the webpages. To verify the claims, we also look into the technologies used in the creation of these webpages and how similar these webpages are to each other when we use the builtwith data as a parameter to compute the Jaccard distance.

When visualized with Gephi, the builtwith similarity for britishdragonshop.com looks as can be seen in Figure 5.10. This clearly has more websites which are worth looking into separately but they include all the webpages identified using the vocabulary parameter. This is a strong indication that the vocabulary revealed sister websites in this particular case.



Nodes: 570 (100% Visible)
Edges: 8 (2.38% Visible)
Undirected Graph

Weighted Degree
Range: 0.0 66.6666...
Label Adjust
Speed 1.0
Include Node size ☑
Avg. Weighted Degree 0.234
Arial Bold, 32

Figure 5.10: *Screenshot of Gephi showing pages similar to britishdragonshop.com with builtwith as a parameter.*

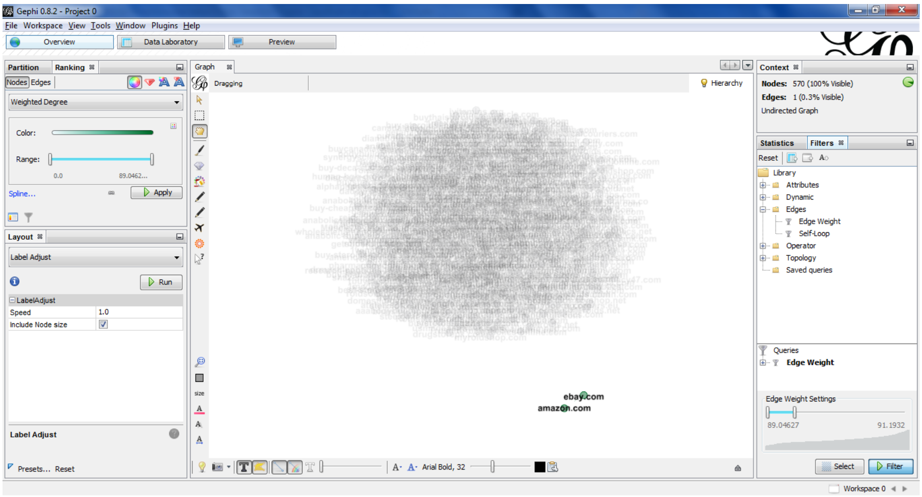
It can thus be said that the vocabulary is indeed a strong metric in determining sister websites that could potentially be run by the same group of people. It may also be possible to identify the most popular and lucrative platforms based on source material originating from this platform being used in multiple other non-related websites that use lets say different technologies to be built.

The reason the second hypothesis is difficult to evaluate would be that the lack of actual ground truth in most of these cases as they happen to be subject of actual criminal investigations.

5.5.2 Replication Unlikely Webpages

The example websites documented as part of this section are likely to not be similar to other websites in the database constructed. The process used to process the data and verify the results is exactly the same as before but the website that it is performed on is different in this section.

If we take the easiest possible example of amazon.com, then we find that the closest looking vocabulary for amazon is that of ebay and even that at close to 0.9, stands significantly different from it as can be seen in the screenshot in Figure 5.11.



Speed 1.0
Include Node size ☑

Nodes: 570 (100% Visible)
Edges: 1 (0.3% Visible)

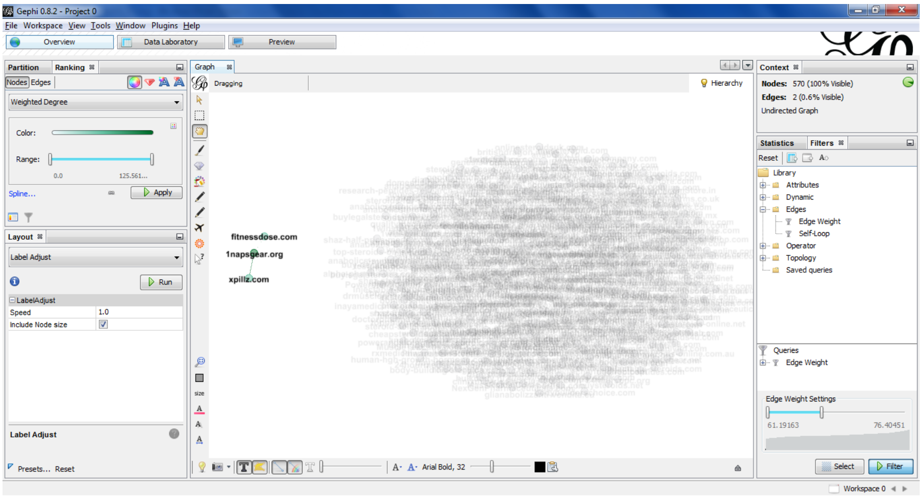
Figure 5.11: *Screenshot of Gephi showing pages similar to Amazon.com*

This shows very clearly that the process can tell dissimilar websites apart quite sharply.

5.5.3 Partial Replication Likely Webpages

As a second alternative, we try to analyze one of the websites from the seed database for similarities where there do not exist many. The website for consideration here is one of the more important ones as described by the investigative team in DBI called 1napsgear.org.

We extracted the HTML_VOCAB Jaccard distance for the website and visualized it using Gephi. It was quickly discovered that there were not many webpages that were similar to this one in the database as the lowest Jdist values were in the order of 0.6-0.7 as can be seen in Figure 5.12, which is quite high but it is a sign that there still were significant similarities.



Gephi 0.8.2 - Project 0

Weighted Degree

Label Adjust

Speed 1.0

Include Node size ☑

Nodes: 570 (100% Visible)

Edges: 2 (0.6% Visible)

Undirected Graph

Edge Weight Settings

61.19163 76.40451

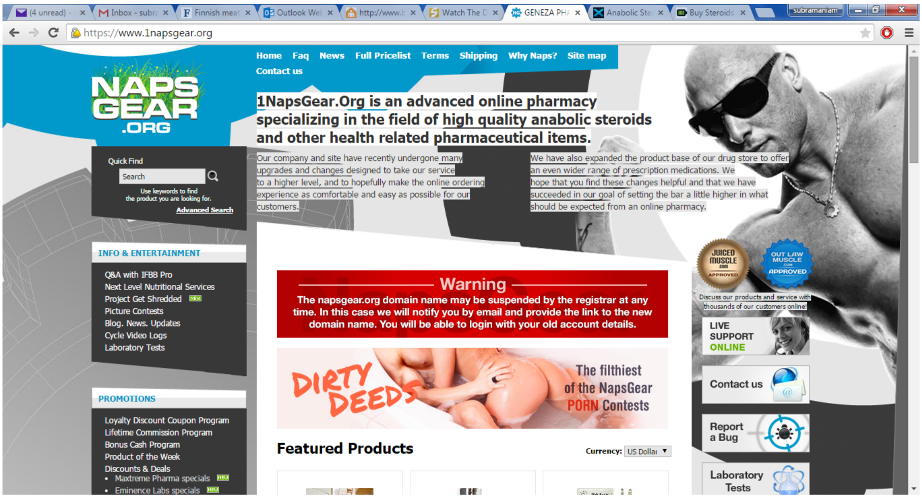
Figure 5.12: *Screenshot of Gephi showing pages similar to 1napsgear.org*

Seeing the homepage of the three websites at first glance does not really reveal any obvious similarities as can be seen the Figures 5.13, 5.14 and 5.15.

But when we take a more comprehensive look at the webpages and their contents we identify striking similarities in layout and content such as the peculiar organization of promotions in all three websites as seen in Figure 5.16. There was also an article quoted in all three websites that spanned a whole paragraph in the middle of the page as can be seen in the screenshots in Figures 5.17, 5.18 and 5.19 in all three of the webpages.

Despite quite a few dissimilarities between the websites, the presence of common content can mean that one of these websites could be the source for the other resellers. This of course needs to be proved through corroboration with other evidence such as new arrivals in the products section and their dates, images and perhaps finally even product batches and shipment details through planned test purchases.

Yet this is definitely considered a good lead to work on and explore.



1NapsGear.Org is an advanced online pharmacy specializing in the field of high quality anabolic steroids and other health related pharmaceutical items.

Our company and site have recently undergone many upgrades and changes designed to take our service to a higher level, and to hopefully make the online ordering experience as comfortable and easy as possible for our customers.

We have also expanded the product base of our drug store to offer an even wider range of prescription medications. We hope that you find these changes helpful and that we have succeeded in our goal of setting the bar a little higher in what should be expected from an online pharmacy.

Warning

The napsgear.org domain name may be suspended by the registrar at any time. In this case we will notify you by email and provide the link to the new domain name. You will be able to login with your old account details.

Discuss our products and service with thousands of our customers online!

Figure 5.13: *Screenshot of 1napsgear.org homepage*

In order to explore this further we look at Gephi graph with builtwith as a parameter to compute the Jaccard distance. The results seen in Figure 5.20 clearly shows that 1napsgear.com is very distant technology wise (least Jaccard distance being in the 0.6-0.7 range). This can be used as data to support the argument that these websites mentioned above are not really sister websites like the ones discussed in Section 5.5.1 but rather just contain borrowed content. Thus we can conclude that the builtwith information can be used to make or break the argument of a website being run by the same criminals.

Hence it is proved that the classification engine can be used to establish automatically the dissimilarities between websites and quickly show websites apart. It is also possible to document that even more interesting details from partial similarities can be used to possibly make more accurate assumptions that lie outside of the scope that webpages can take us to.

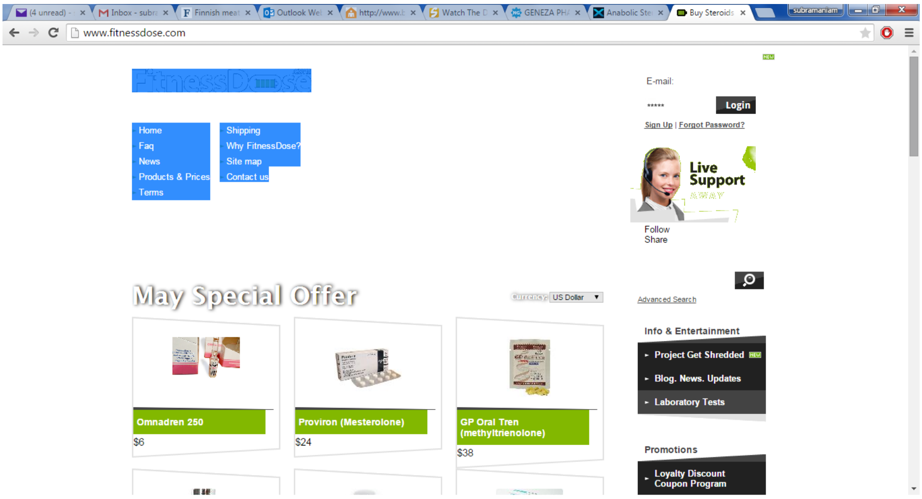
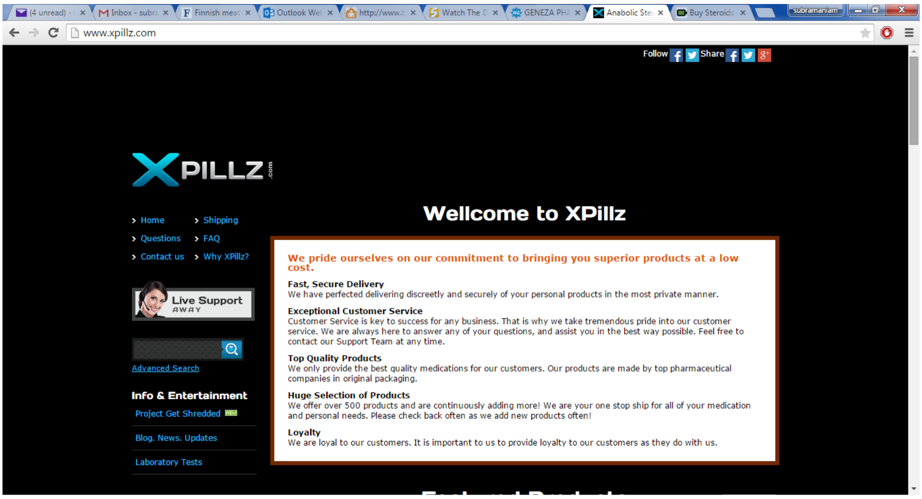


Figure 5.14: *Screenshot of fitnessdoze.com homepage*



Wellcome to XPillz

We pride ourselves on our commitment to bringing you superior products at a low cost.

Fast, Secure Delivery
We have perfected delivering discreetly and securely of your personal products in the most private manner.

Exceptional Customer Service
Customer Service is key to success for any business. That is why we take tremendous pride into our customer service. We are always here to answer any of your questions, and assist you in the best way possible. Feel free to contact our Support Team at any time.

Top Quality Products
We only provide the best quality medications for our customers. Our products are made by top pharmaceutical companies in original packaging.

Huge Selection of Products
We offer over 500 products and are continuously adding more! We are your one stop ship for all of your medication and personal needs. Please check back often as we add new products often!

Loyalty
We are loyal to our customers. It is important to us to provide loyalty to our customers as they do with us.

Figure 5.15: *Screenshot of xpillz.com homepage*

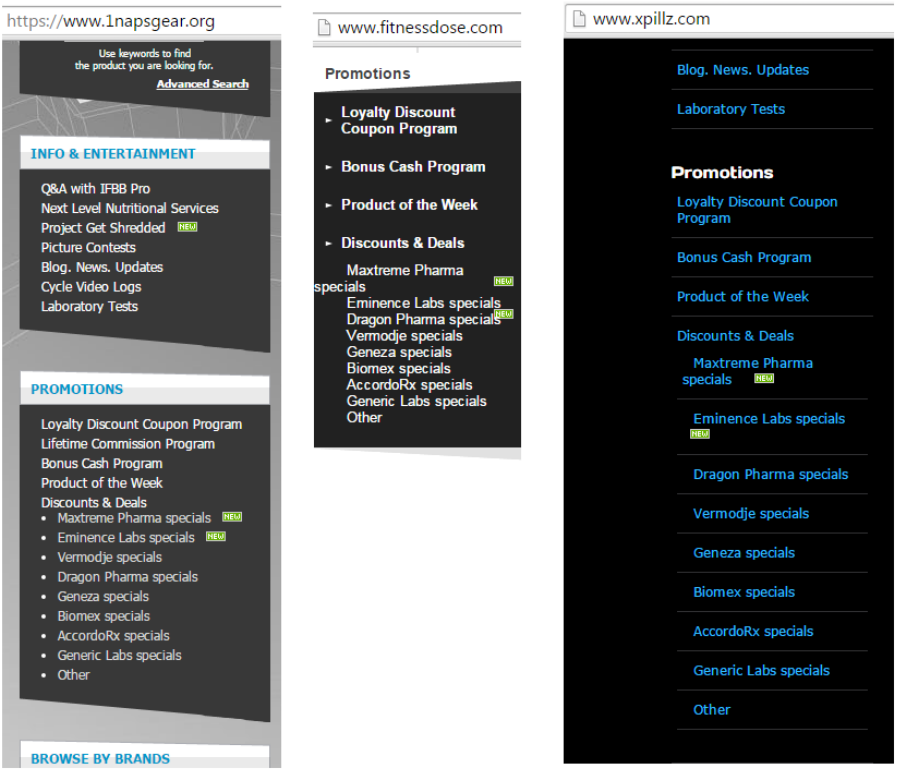
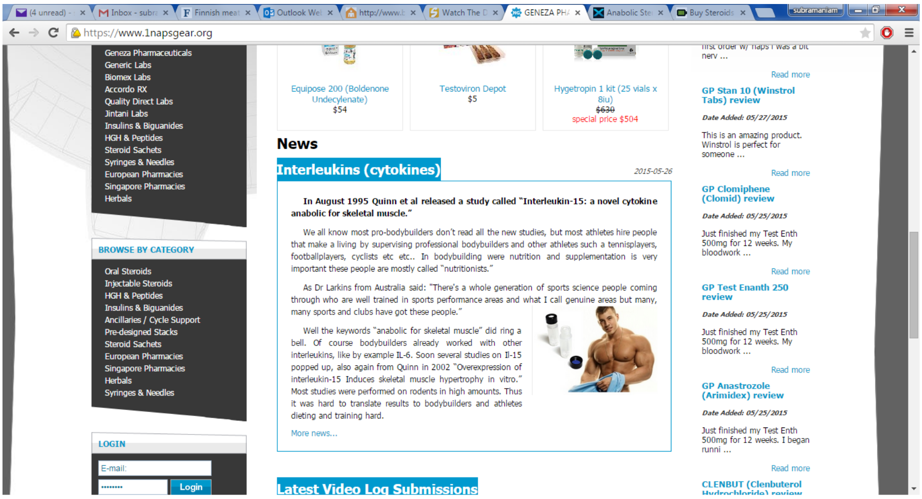


Figure 5.16: *Screenshot of similarities in promotion tab between the 1napsgear.org, xpillz.com and fitnessdose.com*



News

Interleukins (cytokines)

2015-05-26

In August 1995 Quinn et al released a study called "Interleukin-15: a novel cytokine anabolic for skeletal muscle."

We all know most pro-bodybuilders don't read all the new studies, but most athletes hire people that make a living by supervising professional bodybuilders and other athletes such a tennisplayers, footballplayers, cyclists etc etc.. In bodybuilding were nutrition and supplementation is very important these people are mostly called "nutritionists."

As Dr Larkins from Australia said: "There's a whole generation of sports science people coming through who are well trained in sports performance areas and what I call genuine areas but many, many sports and clubs have got these people."

Well the keywords "anabolic for skeletal muscle" did ring a bell. Of course bodybuilders already worked with other interleukins, like by example IL-6. Soon several studies on Il-15 popped up, also again from Quinn in 2002 "Overexpression of interleukin-15 Induces skeletal muscle hypertrophy in vitro." Most studies were performed on rodents in high amounts. Thus it was hard to translate results to bodybuilders and athletes dieting and training hard.

More news...

first order w/ naps I was a bit nerv ...

Read more

GP Stan 10 (Winstrol Tabs) review

Date Added: 05/27/2015

This is an amazing product. Winstrol is perfect for someone ...

Read more

GP Clomiphene (Clomid) review

Date Added: 05/25/2015

Just finished my Test Enth 500mg for 12 weeks. My bloodwork ...

Read more

GP Test Enanth 250 review

Date Added: 05/25/2015

Just finished my Test Enth 500mg for 12 weeks. My bloodwork ...

Read more

GP Anastrozole (Arimidex) review

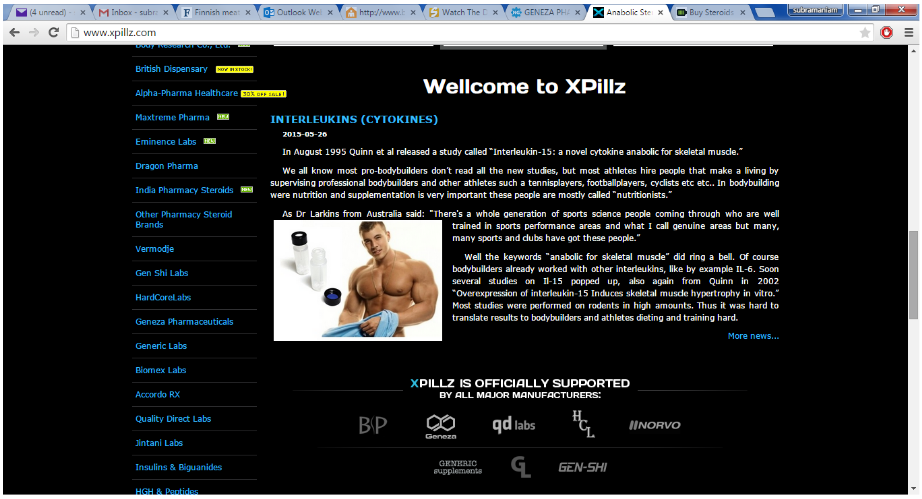
Date Added: 05/25/2015

Just finished my Test Enth 500mg for 12 weeks. I began runni ...

Read more

CLENBUT (Clenbuterol Hydrochloride) review

Figure 5.17: *Screenshot of similar article in 1napsgear.org*



Wellcome to XPillz

INTERLEUKINS (CYTOKINES)

2015-05-26

In August 1995 Quinn et al released a study called "Interleukin-15: a novel cytokine anabolic for skeletal muscle."

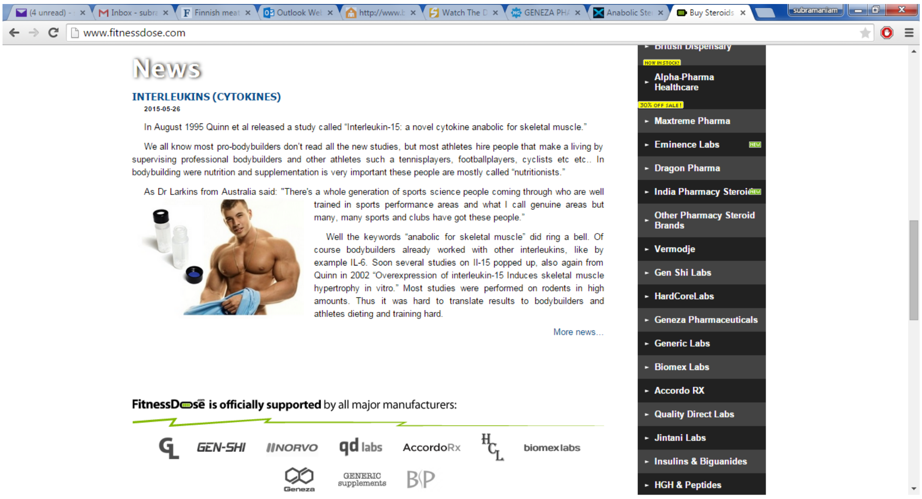
We all know most pro-bodybuilders don't read all the new studies, but most athletes hire people that make a living by supervising professional bodybuilders and other athletes such a tennisplayers, footballplayers, cyclists etc etc.. In bodybuilding were nutrition and supplementation is very important these people are mostly called "nutritionists."

As Dr Larkins from Australia said: "There's a whole generation of sports science people coming through who are well trained in sports performance areas and what I call genuine areas but many, many sports and clubs have got these people."

Well the keywords "anabolic for skeletal muscle" did ring a bell. Of course bodybuilders already worked with other interleukins, like by example IL-6. Soon several studies on Il-15 popped up, also again from Quinn in 2002 "Overexpression of interleukin-15 Induces skeletal muscle hypertrophy in vitro." Most studies were performed on rodents in high amounts. Thus it was hard to translate results to bodybuilders and athletes dieting and training hard.

More news...

Figure 5.18: *Screenshot of similar article in xpillz.org*



News

INTERLEUKINS (CYTOKINES)

2015-05-26

In August 1995 Quinn et al released a study called "Interleukin-15: a novel cytokine anabolic for skeletal muscle."

We all know most pro-bodybuilders don't read all the new studies, but most athletes hire people that make a living by supervising professional bodybuilders and other athletes such a tennisplayers, footballplayers, cyclists etc etc.. In bodybuilding were nutrition and supplementation is very important these people are mostly called "nutritionists."

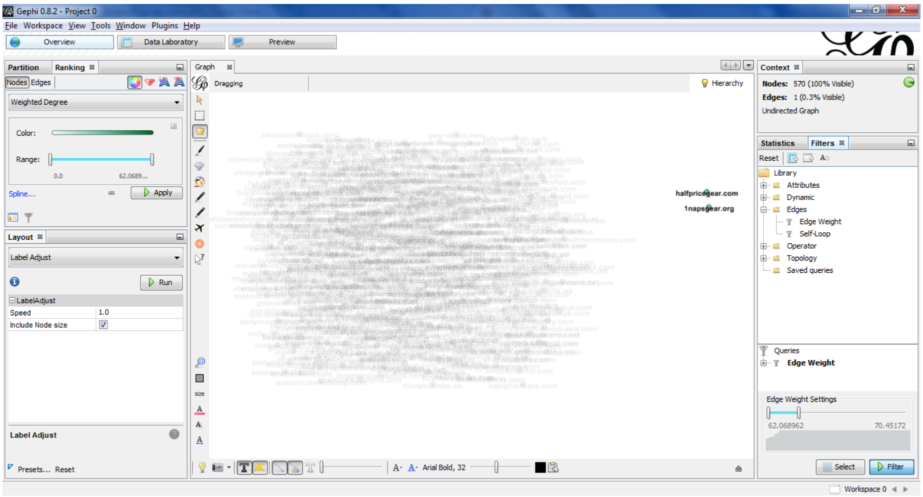
As Dr Larkins from Australia said: "There's a whole generation of sports science people coming through who are well trained in sports performance areas and what I call genuine areas but many, many sports and clubs have got these people."

Well the keywords "anabolic for skeletal muscle" did ring a bell. Of course bodybuilders already worked with other interleukins, like by example IL-6. Soon several studies on Il-15 popped up, also again from Quinn in 2002 "Overexpression of interleukin-15 Induces skeletal muscle hypertrophy in vitro." Most studies were performed on rodents in high amounts. Thus it was hard to translate results to bodybuilders and athletes dieting and training hard.

More news...

FitnessDose is officially supported by all major manufacturers:

Figure 5.19: *Screenshot of similar article in fitnessdose.com*



Weighted Degree

Color:

Range: 0.0 62.0689...

Label Adjust

LabelAdjust

Speed 1.0

Include Node size ☑

halfpricegear.com

1napsgear.org

Nodes: 570 (100% Visible)

Edges: 1 (0.3% Visible)

Undirected Graph

Edge Weight Settings

62.068962 70.45172

Arial Bold, 32

Figure 5.20: *Screenshot of showing pages similar to 1napsgear.com with builtwith as a parameter.*

Chapter 6

Non Functional Requirements

6.1 Introduction

This chapter deals with the detailed description of the non functional requirements of the tool, its purpose and how the tool handles the feature in its implementation to achieve them.

6.2 Architecture and Maintainability

This section documents the software architectural decisions made to keep the tool robust and maintainable. The focus has also been on maintaining standard conventions that will be detailed so that researchers that need to customize and add features to the tool can familiarize themselves with the code to get started quickly.

The entire project is organized into three major modules.

1. Database schema and PL/SQL
2. Processing Component
3. UI Component

6.2.1 Database Schema and PL/SQL

In this section we discuss the schema used by tool and how the various nonfunctional requirements expected of the tool are met from a database perspective.

Normalization is the process of grouping data in a database in such a way that the design forms a clearly understandable, redundancy free(or minimising) data structure with optimal performance [12]. The schema design for the project takes this into account and the database is normalized wherever neccesary to reduce redundancy and keep the design simple.

Care has been taken in steps to reducing redundancy and providing a standard relational database model that conforms to the naming conventions and industry standard normalization practices.

The ER diagram of the database schema can be found in the figure .

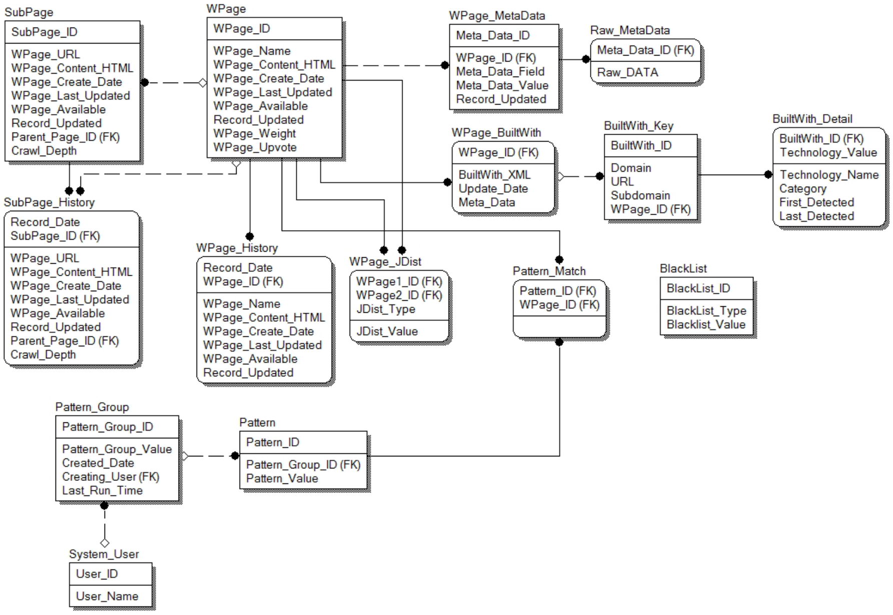


Figure 6.1: *Diagram illustrating the data model of the DB used in the tool.*

The central table used to store the actual content of the webpages homepage and its user voting data is WPAGE. All pages can be uniquely identified by their WPAGE_ID value and the domain is specified as the WPAGE_NAME value.

All subpages that are crawled to from each valid WPAGE is stored in the SUBPAGE table. If a newly domain is newly discovered by crawling, then it needs to be moved manually to the WPAGE table.

The _HISTORY tables are used to store the history values from the SUBPAGE and WPAGE tables when an update is issued on the HTML contents of either. This is achieved through PL/SQL procedures that monitor update activities on the respective base tables.

WPAGE_METADATA table is used to store all the data that is gathered about each WPAGE be it registrar, registrant information, or data that is mined from the actual webpages themselves such as the phrases, tag counts etc. If there is a data value that is larger than can be handled by VARCHAR then it is stored as a CLOB in the RAW_METADATA table.

WPAGE_BUILTWITH, BUILTWITH_KEY and BUILTWITH_DETAIL are the three tables used to store the raw JSON, and the parsed information for the data from the builtwith service for each WPAGE.

The BLACKLIST table contains the information about the keywords that are blacklisted for considerations in the JDist as mentioned in section 4.4.

WPAGE_JDIST table contains the information about Jaccard distances between all pairs of webpages for the various metrics discussed.

PATTERN_GROUP and PATTERN contain the patterns and pattern groups defined by specific users. When the pattern matcher is run, the webpages are parsed and the results of pattern matches are entered into the PATTERN_MATCH table in the schema.

6.2.2 Processing Component

This section talks about the major classes used in the Processing module which caters to the core of the Identification, Discovery and Classification Engines. The class diagrams for the various modules can be seen in figure 6.2 and 6.3.

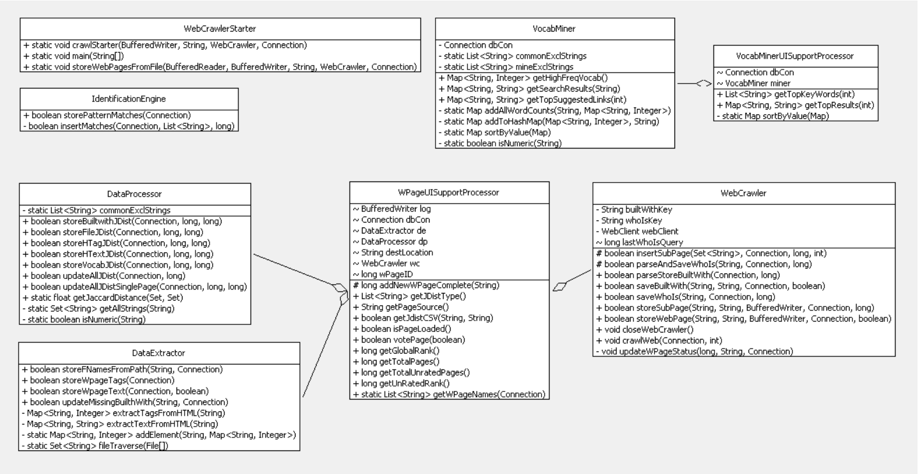

WebCrawlerStarter
+ static void crawlStarter(BufferedWriter, String, WebCrawler, Connection)
+ static void main(String[])
+ static void storeWebPagesFromFile(BufferedReader, BufferedWriter, String, WebCrawler, Connection)
IdentificationEngine
+ boolean storePatternMatches(Connection)
- boolean insertMatches(Connection, List<String>, long)
VocabMiner
- Connection dbCon
- static List<String> commonExclStrings
- static List<String> mineExclStrings
+ Map<String, Integer> getHighFreqVocab()
+ Map<String, String> getSearchResults(String)
+ Map<String, String> getTopSuggestedLinks(int)
- static Map addAllWordCounts(String, Map<String, Integer>)
- static Map addToHashMap(Map<String, Integer>, String)
- static Map sortByValue(Map)
- static boolean isNumeric(String)
VocabMinerUISupportProcessor
~ Connection dbCon
~ VocabMiner miner
+ List<String> getTopKeyWords(int)
+ Map<String, String> getTopResults(int)
- static Map sortByValue(Map)
DataProcessor
- static List<String> commonExclStrings
+ boolean storeBuiltwithJDist(Connection, long, long)
+ boolean storeFileJDist(Connection, long, long)
+ boolean storeHTagJDist(Connection, long, long)
+ boolean storeHTextJDist(Connection, long, long)
+ boolean storeVocabJDist(Connection, long, long)
+ boolean updateAllJDist(Connection, long)
+ boolean updateAllJDistSinglePage(Connection, long, long)
+ static float getJaccardDistance(Set, Set)
- static Set<String> getAllStrings(String)
- static boolean isNumeric(String)
WPageUISupportProcessor
~ BufferedWriter log
~ Connection dbCon
~ DataExtractor de
~ DataProcessor dp
~ String destLocation
~ WebCrawler wc
~ long wPageID
# long addNewWPageComplete(String)
+ List<String> getJDistType()
+ String getPageSource()
+ boolean getJdistCSV(String, String)
+ boolean isPageLoaded()
+ boolean votePage(boolean)
+ long getGlobalRank()
+ long getTotalPages()
+ long getTotalUnratedPages()
+ long getUnRatedRank()
+ static List<String> getWPageNames(Connection)
WebCrawler
- String builtWithKey
- String whoIsKey
- WebClient webClient
~ long lastWhoIsQuery
# boolean insertSubPage(Set<String>, Connection, long, int)
# boolean parseAndSaveWhoIs(String, Connection, long)
+ boolean parseStoreBuiltWith(Connection, long)
+ boolean saveBuiltWith(String, String, Connection, boolean)
+ boolean saveWhoIs(String, Connection, long)
+ boolean storeSubPage(String, String, BufferedWriter, Connection, long)
+ boolean storeWebPage(String, String, BufferedWriter, Connection, boolean)
+ void closeWebCrawler()
+ void crawlWeb(Connection, int)
- void updateWPageStatus(long, String, Connection)
DataExtractor
+ boolean storeFNamesFromPath(String, Connection)
+ boolean storeWpageTags(Connection)
+ boolean storeWpageText(Connection, boolean)
+ boolean updateMissingBuilthWith(String, Connection)
- Map<String, Integer> extractTagsFromHTML(String)
- Map<String, String> extractTextFromHTML(String)
- static Map<String, Integer> addElement(String, Map<String, Integer>)
- static Set<String> fileTraverse(File[])

Figure 6.2: *Part 1 of the class diagram of the Processing module.*

The key elements from the Discovery engine namely the web crawler and the vocabulary miner are placed in the WebCrawler, WebCrawlerStarter and VocabMiner classes

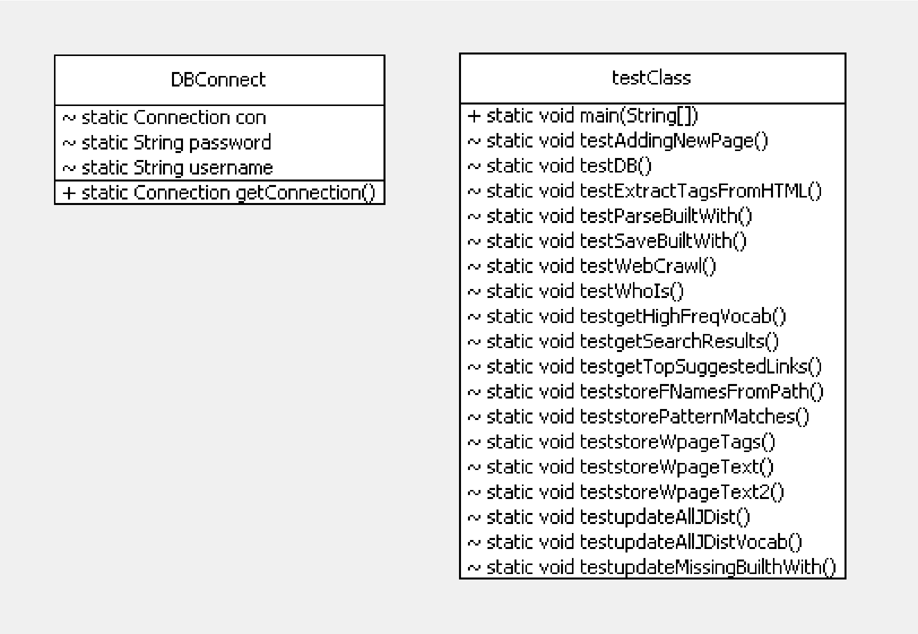

DBConnect
~ static Connection con
~ static String password
~ static String username
+ static Connection getConnection()
testClass
+ static void main(String[])
~ static void testAddingNewPage()
~ static void testDB()
~ static void testExtractTagsFromHTML()
~ static void testParseBuiltWith()
~ static void testSaveBuiltWith()
~ static void testWebCrawl()
~ static void testWhoIs()
~ static void testgetHighFreqVocab()
~ static void testgetSearchResults()
~ static void testgetTopSuggestedLinks()
~ static void teststoreFNamesFromPath()
~ static void teststorePatternMatches()
~ static void teststoreWpageTags()
~ static void teststoreWpageText()
~ static void teststoreWpageText2()
~ static void testupdateAllJDist()
~ static void testupdateAllJDistVocab()
~ static void testupdateMissingBuilthWith()

Figure 6.3: *Part 2 of the class diagram of the Processing module.*

The WebCrawler class has methods that help the user to download a given page, download all the metadata and then try to crawl the webpage for subpages and download them too. The WebCrawlerStarter is a class that can accommodate for mass gathering of data from a list of websites from a file.

All methods take a DBConnection and log file parameter to write/read data from and to and write logs. The only methods that are exposed in each class are the ones that would be used from outside the class.

The DataExtractor class is used to extract data from the HTML pages, such as tag counts, phrases, vocabulary information etc., in the database and update the corresponding tables in the datamodel with the appropriate data.

The DataProcessor class is used to calculate the Jaccard distance for each corresponding measure and store the values in the WPAGE_JDIST table. It offers methods to bulk update/refresh the Jdist for all pairs of websites or for a single website of choice. It also offers the option to chose a specific metric to recalculate as well. The IdentificationEngine class provides methods for user defined pattern matching.

The three of these in effect contribute to the functional features of the Discovery, Identification and Classification engines.

The DBConnect class provides the single source of DBConnection through a static method for the entire tool. This is ensure that changing database connection details happens in only one place either through a configuration file or through direct user input.

The testClass is a unit testing class with static methods to test every major functionality with test data in the tool. It is not exposed in the UI anywhere and must be run from its main method using an IDE. These are the final two classes that can be seen in the class diagram in figure 6.3.

6.2.3 UI Component

Figure 6.4 shows the class diagram for the UI of the tool in question.

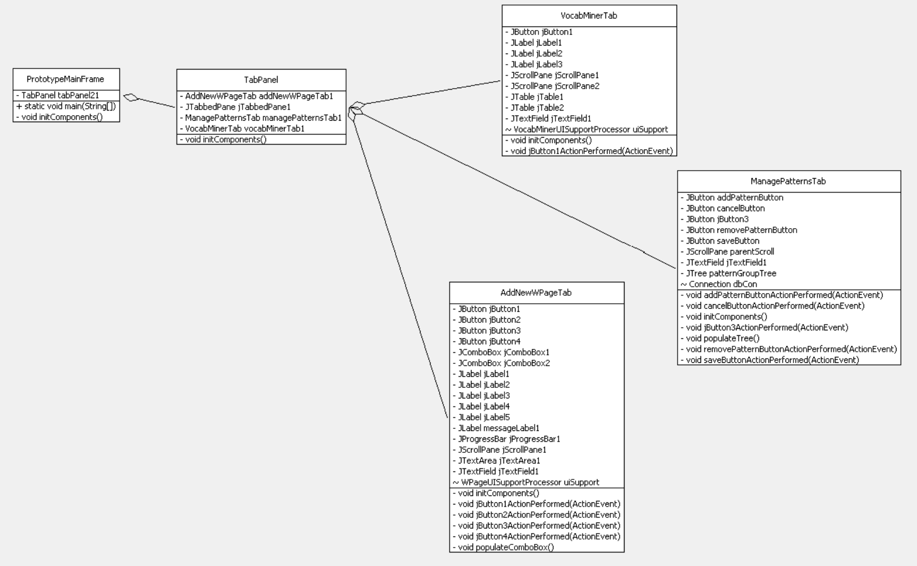

PrototypeMainFrame
- TabPanel tabPanel21
+ static void main(String[])
- void initComponents()
TabPanel
- AddNewWPageTab addNewWPageTab1
- JTabbedPane jTabbedPane1
- ManagePatternsTab managePatternsTab1
- VocabMinerTab vocabMinerTab1
- void initComponents()
VocabMinerTab
- JButton jButton1
- JLabel jLabel1
- JLabel jLabel2
- JLabel jLabel3
- JScrollPane jScrollPane1
- JScrollPane jScrollPane2
- JTable jTable1
- JTable jTable2
- JTextField jTextField1
~ VocabMinerUISupportProcessor uiSupport
- void initComponents()
- void jButton1ActionPerformed(ActionEvent)
ManagePatternsTab
- JButton addPatternButton
- JButton cancelButton
- JButton jButton3
- JButton removePatternButton
- JButton saveButton
- JScrollPane parentScroll
- JTextField jTextField1
- JTree patternGroupTree
~ Connection dbCon
- void addPatternButtonActionPerformed(ActionEvent)
- void cancelButtonActionPerformed(ActionEvent)
- void initComponents()
- void jButton3ActionPerformed(ActionEvent)
- void populateTree()
- void removePatternButtonActionPerformed(ActionEvent)
- void saveButtonActionPerformed(ActionEvent)
AddNewWPageTab
- JButton jButton1
- JButton jButton2
- JButton jButton3
- JButton jButton4
- JComboBox jComboBox1
- JComboBox jComboBox2
- JLabel jLabel1
- JLabel jLabel2
- JLabel jLabel3
- JLabel jLabel4
- JLabel jLabel5
- JLabel messageLabel1
- JProgressBar jProgressBar1
- JScrollPane jScrollPane1
- JTextArea jTextArea1
- JTextField jTextField1
~ WPageUISupportProcessor uiSupport
- void initComponents()
- void jButton1ActionPerformed(ActionEvent)
- void jButton2ActionPerformed(ActionEvent)
- void jButton3ActionPerformed(ActionEvent)
- void jButton4ActionPerformed(ActionEvent)
- void populateComboBox()

Figure 6.4: *Class diagram of the UI Component.*

The PrototypeMainFrame class is the runnable class of the tool which contains the entire panel that is rendered in the UI. The TabPanel is the panel with the tabbed layout for holding each of the actual tabs in the UI.

The VocabMinerTab, ManagePatternsTab and AddNewWPageTag are all panels that correspond to an individual tab in the UI. They contain all the actual elements in the interface and their properties. They use UISupport classes in order to achieve business functionality and merely serve as shapers of data rather than the actual source of data.

To enable operation with the UI, each of the UI feature set is provided with a UISupport class from the Processing Component. These are the WPageUISupportProcessor and VocabMinerUISupportProcessor respectively that can be seen the in figure 6.2.

These classes make use of the actual processing component classes described above, access to the database for other elements that the UI needs and provides them via methods that are relevant to business needs in the UI.

Hence for all UI related enhancements, the change only goes into using the DataProcessor, DataExtractor and WebCrawler intelligently in the corresponding UISupport processor that deals with the tab in question.

6.3 Anonymization

The objective of this section is to help deal with mechanisms to conceal our identity during transactions with potential criminal websites through anonymization. Anonymization of source traffic is a key concern for all applications dealing with criminal investigation.

6.3.1 Background and Purpose

Anonymization is a key concern in investigative tools used by law enforcement for a number of reasons. Owing to better awareness and information sources available to criminals it is becoming common for high tech criminals to identify, track and outmaneuver law enforcement in recent times.

It is very easy for a malicious website to try identify a user either by tracking their IP address or through browser/client characteristic based signatures and spot the common law machines and networks that law enforcements uses. Once this is done, it is easier still for the website to simply use this as a lookup before catering to a request and load a completely different page if necessary with no harmful content in them, to throw off the tool and investigators documenting these websites.

Thus if steps to anonymize our traffic are not taken, then we run the risk of detection and evasion by criminals.

6.3.2 Approaches to Anonymity

To come up with an effective solution to anonymity for the tools traffic, we considered the following options.

Virtual Private Network: The use of a paid VPN service can mean that the traffic is always routed through the standard private network used by the service provider and thus tracking an IP address would always lead to the gateway through which the traffic was routed [9]. This way anonymity can be achieved.

The Onion Router: TOR is a modern powerful tool to ensure anonymity over the Internet. It ensures traffic between all member nodes is passed through a random set of other participating nodes with layers of encryption to ensure obscuring information with each layer of the traffics flow and ensuring end to end anonymity [10]. The Firefox engine based TOR browser has evolved much from its original form and become a stable player in the industry. Taking this up would be an interesting and inexpensive way to provide anonymity.

Dynamic IP Address: If a DHCP line is used to connect to the Internet, we can be sure that the IP address assigned for each login is different from a pool of various addresses from the ISP. This can ensure reasonable levels of anonymity to the end user. But this might not be necessarily possible for a server running on investigative office which predominantly use static addresses for various other reasons. Also since IP and MAC address mappings might be compromised by ISP or other serivice providers, we might choose to go for MAC address spoofing on a peer level to achieve similar anonymity [18].

The key concern for the first two methods is a problem that is difficult to solve namely Captcha. We discuss this in detail at 6.5.1 under the other issues section shortly.

6.4 Performance

In this section we deal with how performance goals in each major component are met from a theoretical and practical perspective to support volume through scalability.

Key concerns, assumptions and bottlenecks are described for each category of features as well.

With a tool such as the one developed for this project, the rote tasks of downloading webpages, parsing information from JSON and storing in the database are fairly monotonous and leave little room for optimization. But scalability is of key concern in these tools as the number of webpages grows, each computation can become increasingly more complex.

This is limited through a few very key optimizations.

6.4.1 Jaccard Distance Computation

One of the key expensive tasks to be performed is that of the Jaccard distance computation between all pairs of websites. This is expensive because as the database grows it takes more and more time to achieve.

If the computation was done for the entire database then it is an $O(n\check{s})$ operation. But we ensure that the operation occurs incrementally only when a new webpage is added. Thus realistically the complexity is always $O(n)$ which is reasonably good.

yet, if the database grows to a large number of websites and the types of measures for which the Jaccard distance needs to be measured also increase, then the operation is going to become more and more expensive. The optimization here was to ensure that only the upvoted database entries were used in this calculation. It is also a consideration for us to only act on and find similarities to websites that are of interest to us. This can help us bring down the complexity substantially if coupled with an option to compute the distance with only currently available websites in addition to the already specified criteria.

6.4.2 Vocabulary Based Calculations

Another key concern during the Jaccard distance calculation for the webpage vocabulary was what the vocabulary of a new webpage needed to be compared against.

If each individual websites vocabulary had to be compared to each other websites individual vocabulary that is a very expensive operation since the vocabulary needs to be created every time a comparison needs to take place.

On the other hand if a universal dictionary exists for all webpages, this can be persisted in to the database and as new websites are added to the DB, they can be updated with additional words to the vocabulary.

This approach was not implemented because of time constraints on the ranking mechanisms possibilities as explained in section 4.4.2. But it is highly recommended as a worthwhile enhancement to the code to improve performance during addition of a new website.

6.4.3 General Guidelines to Improved Performance

Under this section we offer universally applicable recommendations that improve performance.

Parallelism on bulk processing: The tool is a single threaded monolithic process which can be improved dramatically by modeling the UISupport to support multithreaded operation.

High Bandwidth connection: If the bandwidth of the Internet connection is improved, it automatically provides a good boost to the tools performance. This is primarily because the biggest bottleneck to performance in crawling and discovery is still the downloading the actual contents including, images and metadata of many large webpages.

Upgrade to professional WhoIs API: The current API has a time limit of one record access per minute. So there is an inbuilt throttle control mechanism that prevents the tool from trying to make more than one service call per minute. This can be avoided if a completely paid professional WhoIs API service is used instead of the present one.

6.5 Other Issues

In this section we discuss topics that do not fit under the main categories specified above but are still relevant and significant to understanding the problem and developing a meaningful solution to addressing it.

6.5.1 Captcha

Captcha is a simple automated test to tell a human apart from a computer during user interaction. These have become very prevalent in the modern Internet as a strong defense against DOS attacks, automated data gathering tools and scrappers. It is a system that acts directly to prevent exactly what the tool does automatically [23].

This was a key concern during the development of the tool as many of the websites didn't directly implement a weak captcha themselves but rather use captcha services from content service providers such as CloudFlare.

Though the percentage of the websites that used complex and professional captcha services was less, it is still to be considered a very real threat against automated scrapping and discovery capabilities of the web crawler and warrants a keen study on its own.

A few preliminary observations and solutions to the problem, which by no means are comprehensive but a sound quick start, are presented below.

The first step in solving captcha related problems is of course for the web crawler to understand and detect a captch page when it encounters one. This is fairly simple as parsing the HTML and scanning for keywords would reveal this quite easily.

Then one of the following methods can be used to solve the problem or answer the question like a human would.

Use of Image Recognition: There are open source tools that allow for processing weak and poorly formed captcha images and reverse engineer the words like a human would to automatically process captcha [15].

Delayed retrieval: Store the webpage for later retrieval and try once more at a later point of time to see if the page doesnt ask for a captcha again. This can be useful in a DHCP scenario as the IP address may have changed in the meanwhile.

Prompt User Intervention: The least desirable but most accurate of these solutions is ofcourse the involvement of an actual human user. The page can later be retrieved for the user to solve the captcha alone and the tool to perform the automated tasks past that point. It would still be a lot more efficient than a completely manual process but is not as seamless as the vision of the tool demands it to be.

Crowdsourced captcha solving: Another approach to solving captchas can be through the use of human intervention but from a crowdsourcing platform [19]. Here when a captcha page is detected, it is merely redirected to the crowdsourced platforms problem poo and a real human would solve the captcha.

Chapter 7

Conclusion and Future work

7.1 Conclusion

In conclusion we can see that a method using vocabulary based analysis in the discovery, identification and classification of webpages works with a reasonable level of accuracy and help deliver interesting and useful results in the context of counterfeit pharmaceutical products.

It was possible to effectively automate the process of data collection using various openly available services and clever coding to utilize them. The results from the tool have been verified to provide a credible ranking of webpages to segregate new webpages as either ones that conform to the upvoted webpages description or the ones that do not.

The results from the tool have also been verified to provide an accurate discovery mechanism based on the top vocabulary searched over common open Internet search providers.

It has also been proved that with the visualization tools and method prescribed in this report that the process of classifying webpages in the database becomes structured, simple and highly efficient.

The mechanism of utilizing user feedback in order to retain the efficiency of the tool and vocabulary based ranking and classification work very well for the described problem.

It is clear that the tool developed helps provide much needed automation to many manual tasks that burden investigators using innovative and scientific methods.

7.2 Future Work

This section outlines a list of avenues for future research in the area and improvements in the tool developed. It also discusses of possible alternate uses for the developed tool and why they could be an important game changer in the process of discovering, identifying and classifying websites.

7.2.1 Conceptual Work

This section outlines a few concepts that due to time constrain were not tested in this project. It also describes few ideas that were generated during the course of the project which might be worth exploring.

Identification Engine:

1. Ranking of webpages can be based on global dictionary instead of individual dictionary matches.
2. Factor negative weights into ranking pages based on pages that are downvoted.
3. Research on combining results from the use of various parameters for computing Jaccard distance and their impact on classification capabilities of the tool.
4. Development of a global ranking and classification parameter that utilizes multiple parameters described in the report.
5. Research use of more accurate metrics like Cosine Co-efficient that take factors such as frequency into account to improve detection and ranking of similar pages.

Discovery Engine:

1. Identify other mechanisms of web discovery.
2. Research the depth to which the efficiency of crawling is good.
3. Research the optimal number of top links and high frequency words that give good "yield".

7.2.2 Tool related work

This section outlines a few tool related tasks that can yield quick benefit to the investigative community that is toiling hard to keep the web clean. They can not only help mature the tool and make it more useful but also aid in proving the correctness of the approach even more decisively.

Identification Engine:

1. Include subpages into contributing for the vocabulary for a particular webpage.
2. Improved detection of webpages already in the database.

3. Provide capabilities of comparing two specific websites with respect to all parameters.
4. Provide means to maintain different databases for different applications.

Discovery Engine:

1. Implement honeypot based link harvesting
2. Improve control over web crawling by selectively crawling internal/external links only
3. Improve control over web crawling by selectively crawling only specified webpages
4. Customizability and control over additional keywords used in the vocab miner search query.

Classification Engine:

1. Create a seamless interface for importing and visualizing data through customized
2. Provide customizable interfaces for adding other parameters for which datasets can be gathered and Jaccard distance can be calculated.

Appendix A

Appendix

A.1 Vocab Miner Results

A.2 Vocabulary Miner Blacklisted keywords

A.3 Web Crawler Results

Table A.1: *Results from the Vocab Miner for top 25 keywords based search*

Title	Link	Validation
Cheap Deals on iPads, Tablets and E-readers | Go Argos	http://www.argos.co.uk/static/Browse/ID72/33007659/c_1 readers%7C33007659.htm	FALSE
Buy Clomid 50mg Online - About Clomid - Online Drugstore - VERY!	http://www.konell.net/buy-clomid-50mg-online	TRUE
Buy Peptides | The Best Peptide Company | USA Made Peptides ...	https://extremepeptides.com/	TRUE
Buy Winstrol (Stanozolol): Buy Steroids online with 48 Hour free UK ...	http://www.steroids-on-line.com/Buy-Winstrol.html	TRUE
Winstrol Pills for Lean Mass & Strength-Buy Online	http://bodybuildingstackshome.com/legal-steroids/winstrol-pills-review/	TRUE
Tablets: Buy Tablets Online at Best Prices in India | Snapdeal	http://www.snapdeal.com/products/mobiles-tablets	FALSE
Order Deca Durabolin Online	http://www.cyanidecode.org/newsletter/products/deca-durabolin.php	TRUE
Kiwiherb Manuka Oil 10ml, Buy Online Australia | Return2Health	http://www.return2health.net/liquid-supplements/manuka-oil-10ml/	TRUE
Buy Nandrolone Legally	http://steroids-legally.com/buy-nandrolone-legally	TRUE
Xcaret - Chichén Itzá Tour | Chichén Itzá México - Xichen | Tour	http://en.xichen.com.mx/xcaret-chichen-itza-tour.php	FALSE
Amazon.com : Frankincense serrata Essential Oil. 10 ml. 100% Pure ...	http://www.amazon.com/Frankincense-serrata-Essential-Undiluted-Therapeutic/dp/B005V4ZOT2	FALSE
Buy clenbuterol online | buy liquid clenbuterol USA	http://www.madisonjamesresearchchems.com/buy-clenbuterol/	TRUE
Deca Durabolin for Sale: Injectable Nandrolone Decanoate Steroid ...	http://www.roidsmall.net/injectable-steroids-sale-509/deca-duraboline-18174.html	TRUE
Stanozolol tablets - buylegitgear.com - safe place to buy steroids ...	http://www.buylegitgear.com/shop/stanozolol-winstrol-tablets.html	TRUE
Buy Peptides Online | Welcome to Genco Peptides	http://www.gencopeptide.com/	TRUE
Sustanon - Steroids Xtreme	http://anabolicsteroid.biz/sustanon/	TRUE
Buy Winstrol Online - Buy Winstrol-V & Stanozolol Pills	http://www.winstrol.net/	TRUE
Buying Prescription Drugs Online Without Getting Burned - Bloomberg	http://www.bloomberg.com/bw/articles/2013-08-02/buying-prescription-drugs-online-without-getting-burned	TRUE
Anavar 10, powerlifters and bodybuilders Like This | Buy Cheap ...	http://worldanabolicsteroid.com/oral-steroids/anavar/	TRUE

Table A.2: *Blacklisted common English Keywords for the Vocabulary Miner*

BLACKLIST_ID	BLACKLIST_TYPE	BLACKLIST_VALUE
97	JDIST_VOCAB	or
98	JDIST_VOCAB	us
99	JDIST_VOCAB	by
100	JDIST_VOCAB	you
101	JDIST_VOCAB	them
121	JDIST_VOCAB	to
122	JDIST_VOCAB	of
123	JDIST_VOCAB	for
124	JDIST_VOCAB	it
125	JDIST_VOCAB	be
126	JDIST_VOCAB	with
127	JDIST_VOCAB	get
23	JDIST_VOCAB	a
24	JDIST_VOCAB	an
152	JDIST_VOCAB	are
153	JDIST_VOCAB	add
154	JDIST_VOCAB	we
155	JDIST_VOCAB	all
156	JDIST_VOCAB	from
157	JDIST_VOCAB	not
158	JDIST_VOCAB	use
46	JDIST_VOCAB	the
159	JDIST_VOCAB	raw_text
160	JDIST_VOCAB	tabs
161	JDIST_VOCAB	labs
162	JDIST_VOCAB	new
163	JDIST_VOCAB	will
176	JDIST_VOCAB	now
177	JDIST_VOCAB	each
178	JDIST_VOCAB	very
67	JDIST_VOCAB	and
68	JDIST_VOCAB	is
69	JDIST_VOCAB	on
70	JDIST_VOCAB	where
71	JDIST_VOCAB	who
72	JDIST_VOCAB	when
73	JDIST_VOCAB	why
74	JDIST_VOCAB	how
75	JDIST_VOCAB	that
76	JDIST_VOCAB	this
77	JDIST_VOCAB	then

Table A.3: *Blacklisted common English Keywords for the Vocabulary Miner Cont.*

BLACKLIST_ID	BLACKLIST_TYPE	BLACKLIST_VALUE
78	JDIST_VOCAB	as
79	JDIST_VOCAB	if
80	JDIST_VOCAB	at
81	JDIST_VOCAB	here
82	JDIST_VOCAB	in
83	JDIST_VOCAB	while
84	JDIST_VOCAB	no
85	JDIST_VOCAB	yes
86	JDIST_VOCAB	:
87	JDIST_VOCAB	-
88	JDIST_VOCAB	;
89	JDIST_VOCAB	?
90	JDIST_VOCAB	&
91	JDIST_VOCAB	/
92	JDIST_VOCAB	\
93	JDIST_VOCAB	my
94	JDIST_VOCAB	our
95	JDIST_VOCAB	your
96	JDIST_VOCAB	their
179	JDIST_VOCAB	details
180	JDIST_VOCAB	has
181	JDIST_VOCAB	see
182	JDIST_VOCAB	used
255	JDIST_VOCAB	there
256	JDIST_VOCAB	these
257	JDIST_VOCAB	its
258	JDIST_VOCAB	want
259	JDIST_VOCAB	have
260	JDIST_VOCAB	com
261	JDIST_VOCAB	any
262	JDIST_VOCAB	so
263	JDIST_VOCAB	may
264	JDIST_VOCAB	they
265	JDIST_VOCAB	should
266	JDIST_VOCAB	such
267	JDIST_VOCAB	can
268	JDIST_VOCAB	most
279	JDIST_VOCAB	do
288	JDIST_VOCAB	which
289	JDIST_VOCAB	many
290	JDIST_VOCAB	log
291	JDIST_VOCAB	also

Table A.4: *Blacklisted Business specific Keywords for the Vocabulary Miner*

BLACKLIST_ID	BLACKLIST_TYPE	BLACKLIST_VALUE
128	VOCAB_FREQ	home
129	VOCAB_FREQ	products
130	VOCAB_FREQ	contact
131	VOCAB_FREQ	online
132	VOCAB_FREQ	about
133	VOCAB_FREQ	information
134	VOCAB_FREQ	vial
135	VOCAB_FREQ	cycle
136	VOCAB_FREQ	terms
137	VOCAB_FREQ	special
138	VOCAB_FREQ	loss
139	VOCAB_FREQ	privacy
140	VOCAB_FREQ	item
141	VOCAB_FREQ	wish
142	VOCAB_FREQ	list
143	VOCAB_FREQ	other
144	VOCAB_FREQ	payment
145	VOCAB_FREQ	body
146	VOCAB_FREQ	please
147	VOCAB_FREQ	stack
148	VOCAB_FREQ	login
149	VOCAB_FREQ	password
150	VOCAB_FREQ	offer
151	VOCAB_FREQ	specials
164	VOCAB_FREQ	health
165	VOCAB_FREQ	account
166	VOCAB_FREQ	product
167	VOCAB_FREQ	best
168	VOCAB_FREQ	site
169	VOCAB_FREQ	sale
170	VOCAB_FREQ	total
171	VOCAB_FREQ	buy
172	VOCAB_FREQ	pharma
173	VOCAB_FREQ	order
174	VOCAB_FREQ	shipping
175	VOCAB_FREQ	price
183	VOCAB_FREQ	mg
184	VOCAB_FREQ	ml
185	VOCAB_FREQ	delivery

Table A.5: *Blacklisted Business specific Keywords for the Vocabulary Miner Cont.*

BLACKLIST_ID	BLACKLIST_TYPE	BLACKLIST_VALUE
128	VOCAB_FREQ	home
129	VOCAB_FREQ	products
130	VOCAB_FREQ	contact
131	VOCAB_FREQ	online
132	VOCAB_FREQ	about
133	VOCAB_FREQ	information
134	VOCAB_FREQ	vial
135	VOCAB_FREQ	cycle
136	VOCAB_FREQ	terms
137	VOCAB_FREQ	special
138	VOCAB_FREQ	loss
139	VOCAB_FREQ	privacy
140	VOCAB_FREQ	item
141	VOCAB_FREQ	wish
142	VOCAB_FREQ	list
143	VOCAB_FREQ	other
144	VOCAB_FREQ	payment
145	VOCAB_FREQ	body
146	VOCAB_FREQ	please
147	VOCAB_FREQ	stack
148	VOCAB_FREQ	login
149	VOCAB_FREQ	password
150	VOCAB_FREQ	offer
151	VOCAB_FREQ	specials
164	VOCAB_FREQ	health
165	VOCAB_FREQ	account
166	VOCAB_FREQ	product
167	VOCAB_FREQ	best
168	VOCAB_FREQ	site
169	VOCAB_FREQ	sale
170	VOCAB_FREQ	total
171	VOCAB_FREQ	buy
172	VOCAB_FREQ	pharma
173	VOCAB_FREQ	order
174	VOCAB_FREQ	shipping
175	VOCAB_FREQ	price
183	VOCAB_FREQ	mg
184	VOCAB_FREQ	ml
185	VOCAB_FREQ	delivery
186	VOCAB_FREQ	bulk
187	VOCAB_FREQ	search
188	VOCAB_FREQ	items
189	VOCAB_FREQ	shop
190	VOCAB_FREQ	view
191	VOCAB_FREQ	weight
192	VOCAB_FREQ	read
193	VOCAB_FREQ	cycles

Table A.6: *Blacklisted Business specific Keywords for the Vocabulary Miner Cont.*

BLACKLIST_ID	BLACKLIST_TYPE	BLACKLIST_VALUE
128	VOCAB_FREQ	home
129	VOCAB_FREQ	products
130	VOCAB_FREQ	contact
131	VOCAB_FREQ	online
132	VOCAB_FREQ	about
133	VOCAB_FREQ	information
134	VOCAB_FREQ	vial
135	VOCAB_FREQ	cycle
136	VOCAB_FREQ	terms
137	VOCAB_FREQ	special
138	VOCAB_FREQ	loss
139	VOCAB_FREQ	privacy
140	VOCAB_FREQ	item
141	VOCAB_FREQ	wish
142	VOCAB_FREQ	list
143	VOCAB_FREQ	other
144	VOCAB_FREQ	payment
145	VOCAB_FREQ	body
146	VOCAB_FREQ	please
147	VOCAB_FREQ	stack
148	VOCAB_FREQ	login
149	VOCAB_FREQ	password
150	VOCAB_FREQ	offer
151	VOCAB_FREQ	specials
164	VOCAB_FREQ	health
165	VOCAB_FREQ	account
166	VOCAB_FREQ	product
167	VOCAB_FREQ	best
168	VOCAB_FREQ	site
169	VOCAB_FREQ	sale
170	VOCAB_FREQ	total
171	VOCAB_FREQ	buy
172	VOCAB_FREQ	pharma
173	VOCAB_FREQ	order
174	VOCAB_FREQ	shipping
175	VOCAB_FREQ	price
183	VOCAB_FREQ	mg
184	VOCAB_FREQ	ml
185	VOCAB_FREQ	delivery
186	VOCAB_FREQ	bulk
187	VOCAB_FREQ	search
188	VOCAB_FREQ	items
189	VOCAB_FREQ	shop
190	VOCAB_FREQ	view
191	VOCAB_FREQ	weight
192	VOCAB_FREQ	read
193	VOCAB_FREQ	cycles

General Bibliography

Works Cited in this thesis

[1] C. P. Adams and V. V. Brantner, “Estimating the cost of new drug development: Is it really $802 million?”, *Health Affairs*, volume 25, number 2, pages 420–428, 2006 (cited on page 3).

[2] J. Baker, M. Graham, and B. Davies, “Steroid and prescription medicine abuse in the health and fitness community: A regional study”, *European journal of internal medicine*, volume 17, number 7, pages 479–484, 2006 (cited on page 3).

[3] M. Bastian, S. Heymann, M. Jacomy, *et al.*, “Gephi: An open source software for exploring and manipulating networks.”, *ICWSM*, volume 8, pages 361–362, 2009 (cited on page 44).

[4] R. C. Bird, “Counterfeit drugs: A global consumer perspective”, *Wake Forest Intell. Prop. LJ*, volume 8, page 387, 2007 (cited on page 3).

[5] P. H. Bloch, R. F. Bush, and L. Campbell, “Consumer “accomplices” in product counterfeiting: A demand side investigation”, *Journal of Consumer Marketing*, volume 10, number 4, pages 27–36, 1993 (cited on page 3).

[6] R. Cockburn, P. N. Newton, E. K. Agyarko, D. Akunyili, and N. J. White, “The global threat of counterfeit drugs: Why industry and governments must communicate the dangers”, *PLoS medicine*, volume 2, number 4, page 302, 2005 (cited on page 2).

[7] D. Cohen, M. Lindvall, and P. Costa, “Agile software development”, *DACS SOAR Report*, number 11, 2003 (cited on page 12).

[8] J. M. Drew and T. Moore, “Optimized combined-clustering methods for finding replicated criminal websites”, *EURASIP Journal on Information Security*, volume 2014, number 1, pages 1–13, 2014 (cited on pages 30, 32, 35, 43).

[9] E. Gabber, P. P. Gibbons, Y. Matias, and A. J. Mayer, *System and method for providing anonymous personalized browsing by a proxy system in a network*, US Patent 5,961,593, Oct. 1999 (cited on page 68).

[] E. Gallagher, “Compah documentation”, *User’s Guide and application*, 1999 (cited on page 31).

[10] D. Goldschlag, M. Reed, and P. Syverson, “Onion routing”, *Communications of the ACM*, volume 42, number 2, pages 39–41, 1999 (cited on page 68).

[11] M. Jiffriya, M. Jahan, and R. G. Ragel, “Plagiarism detection on electronic text based assignments using vector space model (iciafs14)”, *ArXiv preprint arXiv:1412.7782*, 2014 (cited on page 31).

[12] H. Lee, “Justifying database normalization: A cost/benefit model”, *Information processing & management*, volume 31, number 1, pages 59–67, 1995 (cited on page 62).

[13] L. Li, “Technology designed to combat fakes in the global supply chain”, *Business Horizons*, volume 56, number 2, pages 167–177, 2013 (cited on page 2).

[14] T. Moore, R. Clayton, and R. Anderson, “The economics of online crime”, *The Journal of Economic Perspectives*, volume 23, number 3, pages 3–20, 2009 (cited on page 42).

[15] G. Mori and J. Malik, “Recognizing objects in adversarial clutter: Breaking a visual captcha”, in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, IEEE, volume 1, 2003, pages I–134 (cited on page 71).

[16] A. Perer and B. Shneiderman, “Balancing systematic and flexible exploration of social networks”, *Visualization and Computer Graphics, IEEE Transactions on*, volume 12, number 5, pages 693–700, 2006 (cited on page 44).

[17] M. Rennie, “Claims for the anabolic effects of growth hormone: A case of the emperor’s new clothes?”, *British journal of sports medicine*, volume 37, number 2, pages 100–105, 2003 (cited on page 2).

[18] A. Sayler, “Network anonymity through “mac swapping””, 2011 (cited on page 68).

[19] E. Schenk and C. Guittard, “Crowdsourcing: What can be outsourced to the crowd, and why”, in *Workshop on Open Source Innovation, Strasbourg, France*, 2009 (cited on page 72).

[20] H. Seifoddini and M. Djassemi, “The production data-based similarity coefficient versus jaccard’s similarity coefficient”, *Computers & industrial engineering*, volume 21, number 1, pages 263–266, 1991 (cited on page 31).

[21] K. Song, J. Min, G. Lee, S. C. Shin, and Y.-S. Kim, “An improvement of plagiarized area detection system using jaccard correlation coefficient distance algorithm”, 2015 (cited on page 31).

[22] T. Sørensen, “{a method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons}”, *Biol. Skr.*, volume 5, pages 1–34, 1948 (cited on page 31).

[23] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford, “Captcha: Using hard ai problems for security”, in *Advances in Cryptology—EUROCRYPT 2003*, Springer, 2003, pages 294–311 (cited on page 71).