2013

Intelligent Web Crawling using Semantic Signatures

Lingaiah Choudary Pinnamaneni
West Virginia University

Intelligent Web Crawling using Semantic Signatures

Lingaiah Choudary Pinnamaneni

Thesis submitted to the

Benjamin M. Statler College of Engineering and Mineral Resources

at West Virginia University

in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

Approved by

Dr. Elaine M. Eschen, Ph.D., Chair

Dr. Alan V. Barnes, Ph.D.

Dr. Arun Ross, Ph.D.

Lane Department of Computer Science and Electrical Engineering
Morgantown, West Virginia
2013

Keywords: web crawler, semantic signature

ABSTRACT

Intelligent Web Crawling using Semantic Signatures

Lingaiah Choudary Pinnamaneni

The quantity of test that is added to the web in the digital form continues to grow and the quest for tools that can process this huge amount of data to retrieve the data of our interest is an ongoing process. Moreover, observing these large volumes of data over a period of time is a tedious task for any human being. Text mining is very helpful in performing these kinds of tasks. Text mining is a process of observing patterns in the text data using sophisticated statistical measures both quantitatively and qualitatively. Using these text mining techniques and the power of the internet and its technologies, we have developed a tool that retrieves documents concerning topics of interest, which utilizes novel and sensitive classification tools.

This thesis presents an intelligent web crawler, named Intel-Crawl. This tool identifies web pages of interest without the user's guidance or monitoring. Documents of interest are logged (by URL or file name). This package uses automatically generated *semantic signatures* to identify documents with content of interest. The tool also produces a vector that is a quantification of a document's content based on the semantic signatures. This provides a rich and sensitive characterization of the document's content. Documents are classified according to content and presented to the user for further analysis and investigation.

Intel-Crawl may be applied to any area of interest. It is likely to be very useful in areas such as law enforcement, intelligence gathering, and monitoring changes in web site contents over time. It is well-suited for scrutinizing the web activity of large collection of web pages pertaining to similar content. The utility of Intel-Crawl is demonstrated in various situations using different parameters and classification techniques.

Acknowledgments

I would like to thank my advisors Dr. Eschen and Dr. Barnes for their valuable support, attention to detail and guidance during my research, and also for making me a part of Discrete Algorithms Research Team. Successful completion of my thesis can be attributed to the guidance given by my advisors. I thank my family especially my parents and beloved one Shalini Koney who played a major role behind my journey of Masters at West Virginia University. I would like to thank Dr. Ross for the support and valuable guidance at hard times. Thanks to my student colleague Ravali Kota for her valuable suggestions, which helped me a lot in achieving any goals regarding this thesis. I would like to thank all my friends and LCSEE Department. A special thanks again to my parents in guiding every step of my career to have a successful life.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. Introduction

We live in a world that is connected via the internet, where data is increasing in huge volumes. Here processing or acquiring the required data from that large heap is a tedious task. The majority of the data is in text format. Text mining was developed for the purpose of extracting content of interest from large piles of data. Text mining is also a process of acquiring high quality text from natural unstructured text [1]. While the internet is a vast storehouse of text data, where both the content and who has posted it can be of interest, the data is highly unstructured, and thus, difficult to navigate. A web crawler uses existing hyperlink structure to access documents. Our tool refines the existing link structure by following only links that lead to documents with content of interest. Here text mining aids the user in discovering documents with the targeted content and the relationships between documents.

There are two major things in achieving this goal. First, we need to preprocess the text data found on the web, since there is no fixed rule or language to publish text on the web; for example, one may use a normal HTML to publish a web document or ASP to publish the similar content. But, we need to extract the data related irrespective of the format used. Second, we need to filter out documents that are not of interest and extract the exact required content from the collected data. Various text mining tools and clustering algorithms are used to analyze the retrieved documents.

In this thesis we develop a web based tool that crawls web pages to extract content of interest using *semantic signatures*. Semantic signatures provide a quantification of targeted content that can capture nuances of a single topic. The semantic signatures are automatically generated, using previously developed software and algorithms, with minimal external input.

Motivation

This idea was generated based on the Automated SSMinT package, which is a powerful and intelligent text mining tool developed by E. Eschen, A. Barnes, R. Kota, U. Para, and S Peddada [2,3,4]. This package has sequence of tools which discover and refine *semantic signatures* in known content documents. These semantic signatures can later be used retrieve documents with similar content from a corpus of unknown content documents and to categorize a collection of documents. In this thesis we extend the use of semantic signatures to the corpus presented by the web. Crawling the web using semantic signatures yields good results in fetching the targeted content instead of lots and lots of unwanted data in less time.

Automated SSMinT package has its applications limited only to the stand alone system in which a fixed corpus of text data is presented to the system for analysis. The WVU Discrete Algorithms Research Team (DART), led by Drs. Eschen and Barnes, is interested in discovering documents with targeted content on the web, and further generating the link relations between these documents. Such tools can be used by law enforcement investigators and intelligence analysts. Further, studies of the data on the web over a period of time can illuminate important shifts content. These are some of the motivations behind developing the Intel-Crawl software package.

Contribution

In this thesis an intelligent web crawler tool that crawls over the web using an initial URL to acquire web documents of interest is designed and implemented. The previously developed Hits Array Generator Tool [2,3,4] is integrated into the web crawler module for the generation of a Hits Array. In this, four modules were designed to perform the following functionalities:

1. Input the URL of a seed web page and download the web page. Retrieve both text and other web links from it.
2. Using the links, download the corresponding web pages. The module crawls the web and repeats the step 1 to retrieve further the text and web links.
3. In each of the set of links, each link is crawled. The text is extracted and analyzed. Semantic signatures present in a document are recorded in a row of the Hits Array. The URLs of documents for which at least one semantic signature is recorded are added to a list.
4. All such document URLs are found in a breadth-first-search of the web link graph rooted at the seed URL. A Hits Array is generated that contains a vector for each document of interest. This vector stores the frequency of each semantic signature in the document.

The resulting Hits Array can be further analyzed to classify the retrieved documents.

Flow of Document

In this thesis document, each chapter is designed as follows:

Chapter 2 describes the background concepts that are required to understand the basic terms that are encountered in the following chapters.

Chapter 3 details the previous work that is used in implementing Intel-Crawl. In this, concepts of Automated SSMinT are explained in brief, including how each tool is implemented and connected to other tools. The inputs and outputs of each tool are described.

Chapter 4 gives the details of the Intel-Crawl Web, its components and the functionality.

Chapter 5 outlines various experiments conducted and their results. Experiments were designed to test the performance of Intel-Crawl and its efficacy in retrieving documents of interest from the web corpus.

Chapter 6 explains the future work of the current project, and includes a conclusion part that gives the final overview of Intel-Crawl, its advantages and disadvantages.

2. Background Concepts

This chapter gives the brief explanation of various concepts, terminology, and previously developed methods that will be used in the sequel.

Term Frequency-Inverse Document Frequency (TF–IDF)

Term Frequency-Inverse Document Frequency [6] is a weight measure that reflects the importance of a word to a document, which is in a collection D of documents. The importance increases proportionally to the number of times the word appears in the document, but is modified to account for the fact that some words are far more common than others. Term frequency denotes the frequency of a word occurring in a document. Inverse Document Frequency is the weight that increases the importance of less frequently occurring in the corpus.

Multiplying TF and IDF will pop out the relevant terms to a document. Priority is given to the term that appears frequently in the document and fewer times in the corpus. The priority words are given higher weight. Thus, it prevents the highly common words from being assigned high weights. Applications of TF-IDF are used in search engines and information retrieval systems due to its simplicity and effectiveness.

Mathematically, the equations involved in calculating term frequency and inverse document frequency are as follows [7]:

TF(t, d) = (number of times the term t appears in the document d) / (number words in the document)

IDF(t, d) = $\log\frac{|D|}{|\{d\in D:\ t\in d\}|}$

K-Means Clustering

The K-Means algorithm is an unsupervised clustering algorithm. The algorithm divides M vectors in N dimensions into K clusters, the within-cluster sum of the squares is calculated for each cluster, and the sum of these is minimized [8]. The minimizing criterion is:

$$J = \sum_{j=1}^{K} \sum_{n\in M_j} \left| x_n - \mu_j \right|^2$$

where x_n is a vector representing the nth data point and μ_j is the centroid of the data points in cluster M_j.

The implementation of the algorithm [9] is as follows after assigning data points randomly to K sets:

- Centroid is computed for each set.
- Every point is assigned to the cluster whose centroid is closest to that point.
- Above steps are repeated until the halt criterion is met.

K-Means Clustering has been applied in different real-time applications such as market segmentation, computer vision, geo statistics, astronomy, and some parts of agriculture.

Cosine Similarity Measure

The cosine similarity measure [11] is the cosine angle between two vectors. This helps in finding the orientation of two vectors; i.e., whether they are oriented in the same direction or not. For two vectors X and Y the cosine similarity measure ‘cos θ’ is defined as:

$$\cos\theta = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \times \sqrt{\sum_{i=n}^{n} Y_i^2}}$$

The values near to 1 indicate that two vectors are oriented in a similar direction. The advantage of using this similarity measure is it can be seen as normalizing document length when making content comparisons based on term frequencies.

Singular Value Decomposition

Singular Value Decomposition (SVD) of a matrix M is defined as the factorization of M into U$\sum V^t$, where M is m×n real or complex matrix, U is a m×k real matrix or complex unitary matrix, $\sum$ is a n×n diagonal matrix and V^t is a k×n real or complex unitary matrix, and $k \leq n$ [12]. The diagonal entries of Σ are known as the *singular values* of *M*.

$$M_{[m\times n]} = U_{[m\times k]} \sum\nolimits_{[k\times k]} V^t{}_{[k\times n]}$$

Note that *k* may by less than *n* in the above SVD. *Dimensionality reduction* is done by setting the smaller singular values to 0. The main purpose of dimensionality reduction is to remove the irrelevant dimensions which carry little or no information and contribute to noise that affects analysis.

Breadth First Search Algorithm

Breadth First Search (BFS) is a graph traversal method that visits all the children a node before traversing to any successor of a child node [13]. The algorithm is stated as follows:

Statement: Given an input graph G = (V, E) and source vertex s, from where to begin. BFS systematically explores the edges of G to discover every vertex that is reachable from s. It produces a breadth first search tree with root s that contains all such vertices that are reachable from s. For every vertex v reachable from s, the path in the breadth first tree from s to v corresponds to a shortest path.

Following steps are to be followed for this algorithm to be executed

- Step 1: Initialize all nodes to ready state (status = 1)
- Step 2: Put the starting node in queue and change its status to the waiting state (status = 2)

- Step 3: Repeat step 4 and 5 until queue is empty
- Step 4: Remove the front node n of queue. Process n and change the status of n to the processed state (status = 3)
- Step 5: Add to the rear of the queue all the neighbors of n that are in ready state (status = 1), and change their status to the waiting state (status = 2). [End of the step 3 loop]
- Step 6: Exit

Note, BFS can be applied to both undirected and directed graphs (networks). In crawling web pages we use this algorithm. How we implement the algorithm will be explained in the later chapters.

Literature Review

The basic idea of a web crawler is to crawl over the web and fetch the documents that are related to the user's topic of interest. Dong and Hussain [14] proposed semantic based focused crawling in which they classified the Web Crawlers into two categories: *ontology based* web crawlers and *metadata based* web crawlers. Our intelligent web crawler is similar to the concept of metadata based web crawler. There are some shortcomings of metadata based web crawlers as defined by Dong and Hussain [14]. They are as follows:

- The plain text classifiers don't have enough semantic supports and this causes poor performance.
- The documents may not match the exact concept. This may not meet the user's requirements.

The above shortcomings are addressed in this thesis using *semantic based* crawling. Regarding the semantic support of the text documents, we use *semantic signatures* based on *keyword sets*. For a given document (of unknown content), a collection of vectors are generated that contain data on the frequency, proximity, and order of keywords in the document. Using the same keyword sets, vectors are generated for training documents (of known content). These vectors are clustered to form semantic signatures. The semantic signatures quantify concepts in the training documents. The semantic signatures are stored for use by text mining applications, including filtering and categorization. The second problem is addressed by the power and sensitivity of semantic signatures in capturing the concepts in a text document. Automated Learner Tool, which will be explained more in chapter 3, automatically generates and refines semantic signatures from training documents, thereby providing the user with a collection of semantic signatures for the most important concepts contained in a document.

The system architecture defined by Dong and Hussain is similar to that of our web crawler's architecture with slight variations in the implementations.

Implementations of web crawlers are addressed in various situations such as the vertical portal system defined by Francesconi and Peruginelli [15] and in an e-business information search engine to generate metadata based on the web documents downloaded using the CiteSeer technique as defined by Giles, et al. [16].

3. Automated SSMinT Package

Introduction

This chapter gives the brief overview of the Automated SSMinT package developed by Barnes, Eschen, and Kota [2]. In Automated SSMinT there are various tools: Automated Keyword Tool (AKT), Automated Learner Tool (ALT), Hits Array Generator Tool (HAGT), and Semantic Signature Refinement Tool (SSRT). These tools are developed to extract the semantic signatures of the targeted content without human intervention. The hand-drawn figures in this chapter were drawn by L. Pinnamaneni and also appear with permission in [2].

The process is sequential and the final output is derived in the Semantic Signature Refinement Tool. The flow starts with the Automated Keyword Tool (AKT), which takes as input training documents, window size, window function constant, list of synonyms to be replaced, and the option whether word stemming should be used. AKT outputs keyword sets that are stored in Keyword Descriptor Files (KDFs). Keyword sets are generated based on TF-IDF and forward and backward distances. Then the KDFs are input to the Automated Learner Tool (ALT), which generates semantic signatures using the cosine distance measure and a clustering algorithm. ALT generates document vectors which are refined and clustered; selected clusters are stored in Semantic Signature Descriptor Files (SSDFs). SSDFs are given as input to the Hits Array Generator Tool (HAGT), which generates a Hits Array with semantic signatures as columns and documents of unknown content as rows. Each document row stores the counts of the number of occurrences of the various semantic signatures in that document. The Hits Array is stored in a Hits Array Descriptor File (HADF) and passed to the Semantic Signature Refinement Tool (SSRT), which refines the semantic signatures and also removes the redundant relevant semantic signatures from the group. The total flow of the package is shown in Figure 1, which is adapted from Barnes, Eschen, and Kota [2]. Now we describe each tool in detail.

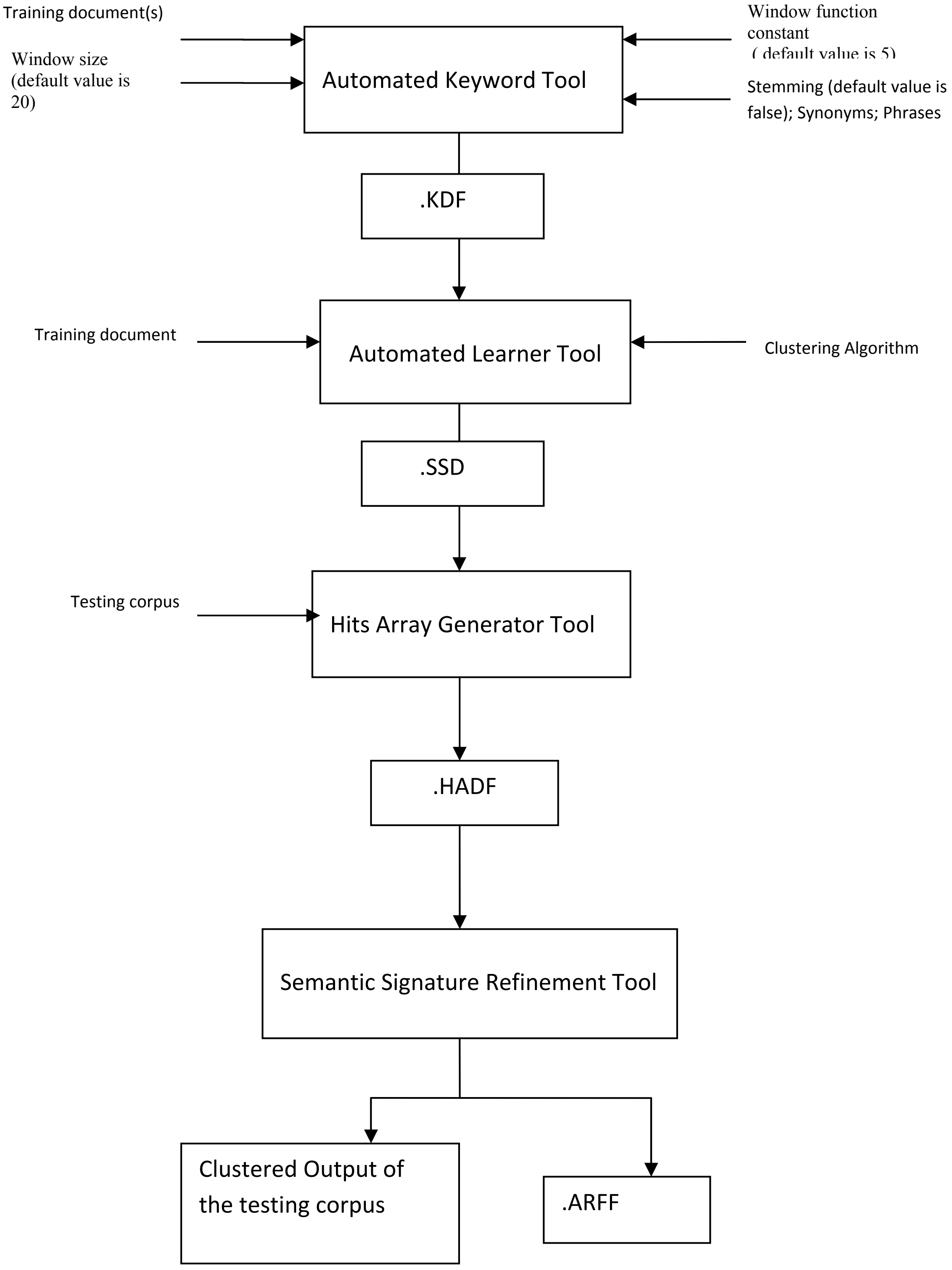
Training document(s)
Window size (default value is 20)
Automated Keyword Tool
Window function constant (default value is 5)
Stemming (default value is false); Synonyms; Phrases
.KDF
Training document
Automated Learner Tool
Clustering Algorithm
.SSD
Testing corpus
Hits Array Generator Tool
.HADF
Semantic Signature Refinement Tool
Clustered Output of the testing corpus
.ARFF

Figure 1. Flow of tools in Automated SSMinT package [2]

Automated Key Word Tool

The Automated Key Word Tool (AKT) [2] was developed for automatic generation of keyword groups without intervention of an analyst. This tool is mainly used in Intel-Crawl in the process of generating a semantic signature library. The training documents along with the window size, list of synonyms to be replaced and stemming options are given as input. First, the tool will check whether the stated documents exists or not; if they exist, it will process them to calculate the term frequencies within each training document and the inverse document frequencies relative to the entire corpus. It generates an ordered list of words that have high importance. These words are used for generation of keyword groups using a window weight calculation technique. From the list of the words generated the first five words are chosen and the weights are calculated for each and every word in a window using both forward and backward distances, from these three strongly correlated words are chosen, while avoiding duplication of words. Then keyword groups are generated with all possible combinations of the chosen words. The keyword groups are written in a Keyword Descriptor File (KDF) and stored in the destination folder.

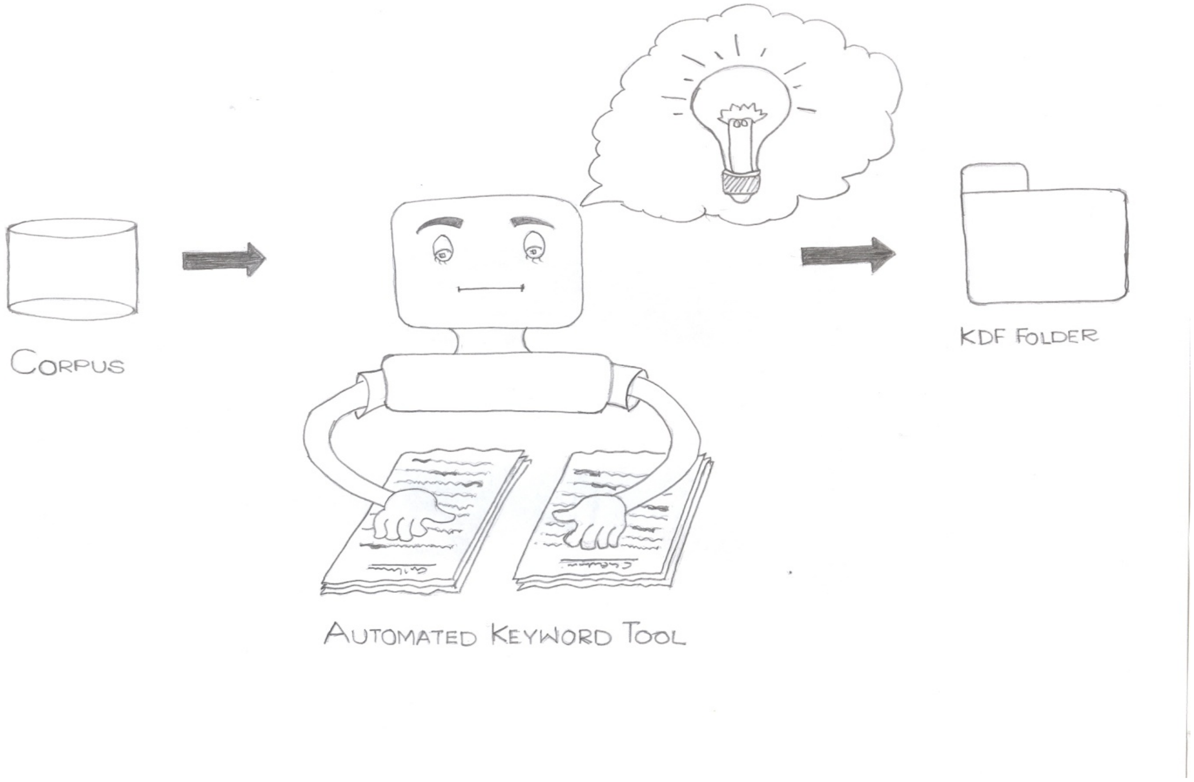

Figure 2. Automated Keyword Tool

Automated Learner Tool

This tool discovers the semantic signatures automatically and they are written in a Semantic Signature Descriptor File (SSDF). Automated Learner Tool takes as input KDFs, training documents, and information regarding the clustering algorithm. First, using the training documents and KDFs, document vectors are generated. Using vector refinement and the clustering algorithm specified semantic signatures realized as clusters of vectors are saved in Semantic Signature Descriptor Files (SSDFs) in XML format with.ssd extension.

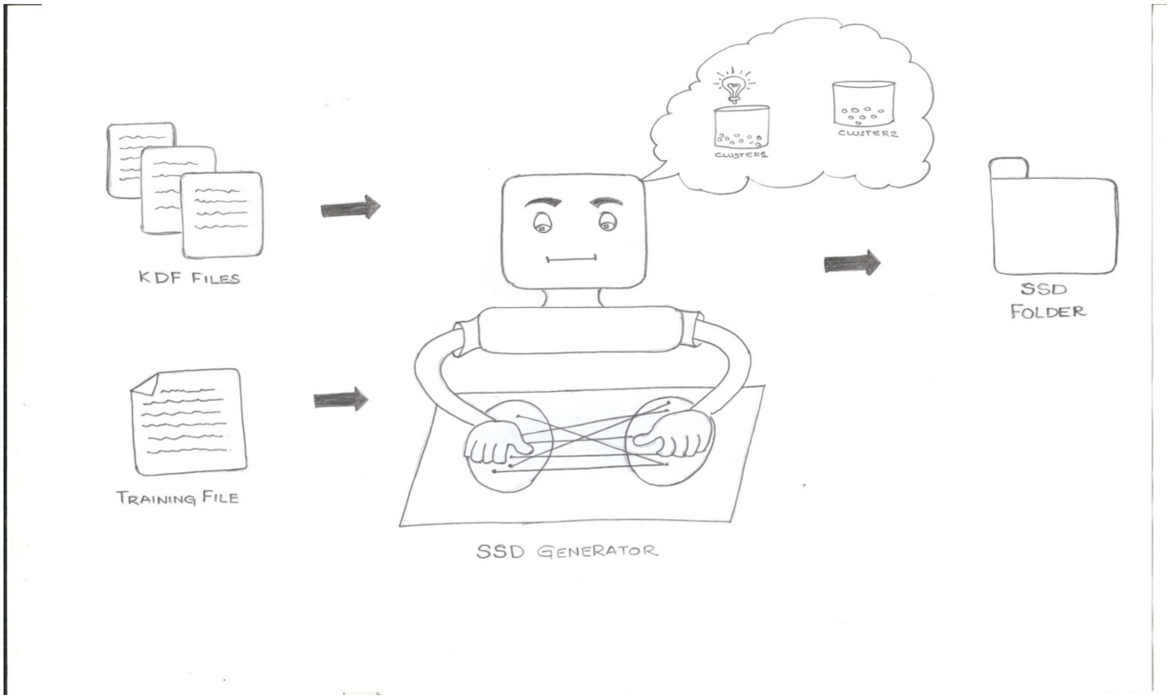

Figure 3. Automated Learner Tool

Hits Array Generator Tool

This tool generates a Hits Array, which is a document versus semantic signature matrix. It takes as input the Semantic Signature Descriptor Files and testing documents. Then the tool performs the folder existence check and then semantic signature hits are calculated. The cluster definition for semantic signatures is fixed as CD2 [3] in Intel-Crawl; a *hit* in a [document, semantic signature] array cell is generated if and only if the cosine similarity between a testing document vector and the centroid of the cluster is less than the angle obtained from the Semantic Signature Descriptor File. The semantic signature (cluster) hits are stored in a Hits Array Descriptor File (HADF).

The hits calculation performs various roles in Intel-Crawl because HAGT is used for both generation of a semantic signature library and for crawling of web pages. The Hits Array Descriptor File, which is written in XML format with .hadf extension, contains the list of semantic signatures used, the link that was crawled, and the hit values.

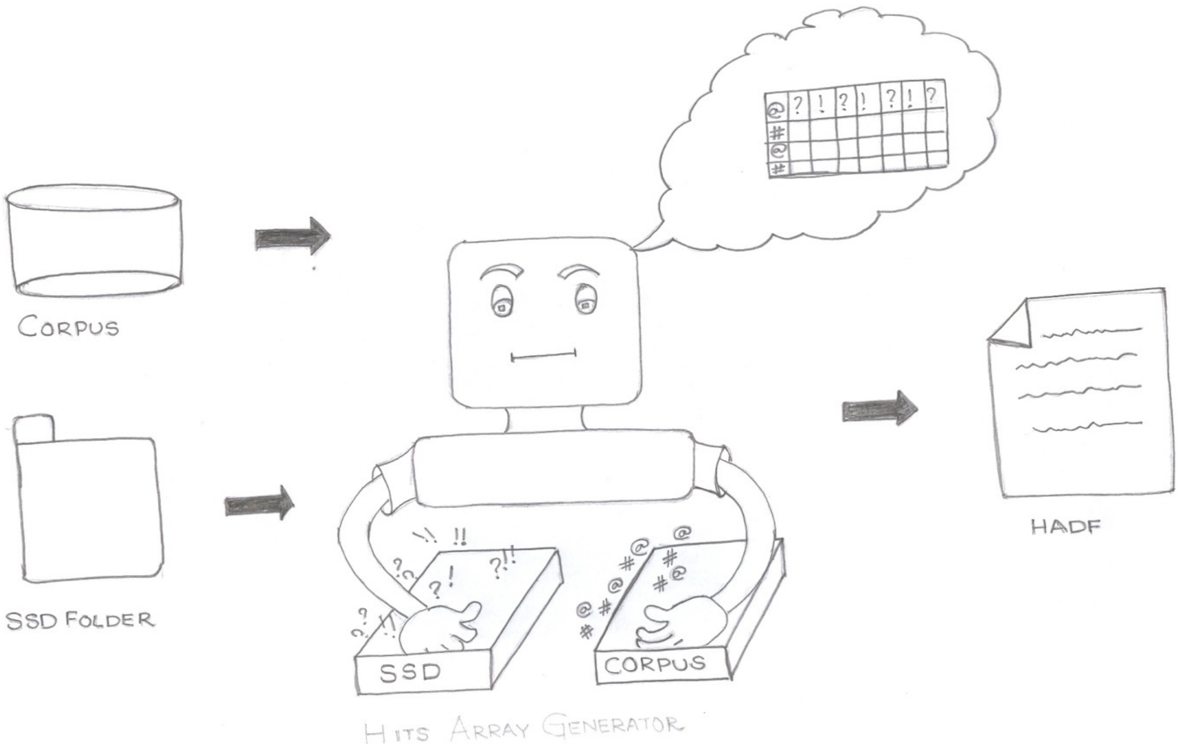

Figure 4. Hits Array Generation Tool

Semantic Signature Refinement Tool

This tool is used to refine the semantic signatures and also for pruning relevant redundant semantic signatures. This tool takes as input the Hits Array Descriptor File (HADF) and groups semantic signatures based on their relevancy; the semantic signature column vectors are clustered using the K-Means clustering algorithm. It uses the Singular Value Decomposition (SVD) technique to identify strong semantic signatures and then prune some redundant relevant semantic signatures. Using SSRT the user can perform different tasks such as refining automatically generated Semantic signatures, clustering document row vectors using K-Means, and converting the .hadf file to .arff file, which is an Attribute Relational File Format used by WEKA.

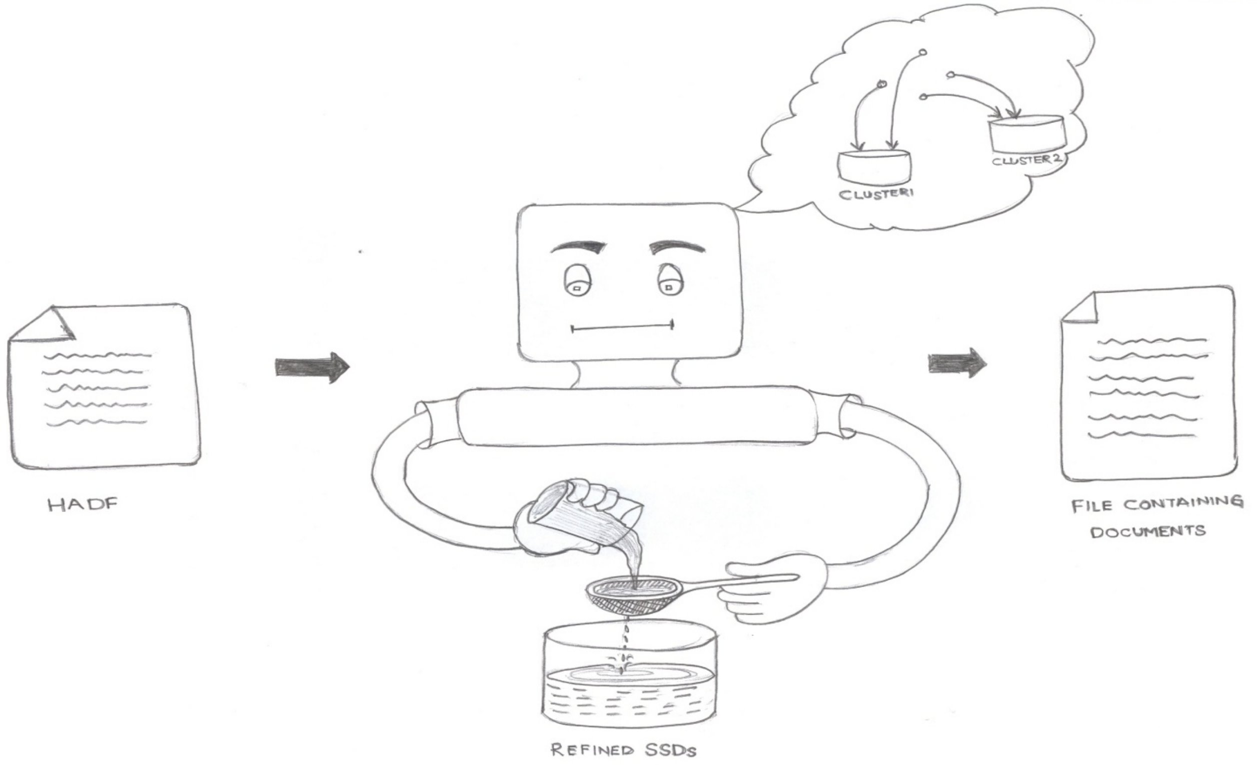

Figure 5. Semantic Signature Refinement Tool

Implementation of Automated SSMinT Package in Intel-Crawl

Automated SSMinT is used for two different applications: generation of a semantic signature library and crawling web pages.

In generation of a semantic signature library, first the training and testing URLs are collected manually and input to the crawler unit which fetches the textual information and other URLs from them. These documents are used as an input for Automated SSMinT. The collection of testing documents is used to approximate the corpus, which in reality is the WorldWideWeb, for the purpose of calculating IDFs. First the folder of training documents and folder of testing documents are prepared. Then they are given as input to Automated Keyword Tool (AKT). Thus, the keyword groups are generated and stored in KDFs. The KDFs along with the training documents are given as input to Automated Learner Tool (ALT). ALT outputs semantic signatures which are stored in Semantic Signature Descriptor Files. These SSDFs along with the testing documents are given as input to Hits Array Generator Tool (HAGT). This tool produces the Hits Array, which is a document versus semantic signature matrix. The matrix is stored in a Hits Array Descriptor File (HADF). The HADF is input to Semantic Signature Refinement Tool (SSRT), which refines the semantic signatures and prunes redundant relevant semantic signatures. The resulting semantics signatures are saved in a library.

The second part of Intel-Crawl is integration of the web crawler with Hits Array Generator Tool for regular crawling of web pages. In this, a seed URL is given as input and then the web content is downloaded and the textual data is extracted. This is passed to the integrated Hits Array Generator Tool (HAGT) to generate a row of the document-semantic signature matrix using the semantic signature library (provided at least one semantic signature hit is found). The current document's links are then crawled. The Breadth First Algorithm is implemented for crawling based on the hits generated each time. A URL and its corresponding chain of links is considered only if the parent URL has at least one hit, else everything will be discarded and the next URL is processed.

4. Intelligent Web Crawler – Intel-Crawl

Motivation

The Automated SSMinT [2] (Semantic Signature Mining Tool) package developed by Barnes, Eschen, and Kota, categorizes or filters the corpus into groups without any help of an analyst based on semantic signatures that are discovered by the tool. This is a generic package that categorizes text documents without a priori definition of classes. It was designed to accept as input a static corpus of documents.

Extending the idea of grouping/filtering a large set of documents in a corpus to the next level, the WVU Discrete Algorithms Research Team (DART), led by Drs. Eschen and Barnes, came up with an idea of developing a web application that takes the basis of Automated SSMinT to the web level in categorizing/filtering a large set of web pages.

The motivation for Intel-Crawl is to develop a tool that observes the web activity of large set of web pages based on the topics of interest. This tool crawls a large set of interrelated and interlinked web pages, discovering pages of interest to an analyst. This helps the analyst in observing a large set of web activity at once without guidance or monitoring by the analyst. Documents of interest are logged (by URL or file name). The log provides analysts with a starting point for further investigation. The tool utilizes semantic signatures generated by Automated SSMinT to determine the content of web pages. It also produces a vector that is a quantification of a document's content based on the semantic signatures. This provides a rich and sensitive characterization of the document's content.

Intel-Crawl may be applied to any area of interest. It is likely to be very useful in areas such as law enforcement, intelligence gathering, and monitoring changes in web site contents over time. It is well-suited for scrutinizing the web activity of large collection of web pages pertaining to similar content.

Overview of Intel-Crawl

The development of the web crawler has three stages which consist of: i) building the training and testing sets, ii) generating a semantic signature library, and iii) crawling the web. Building testing and training sets are the primary steps in development for applying the web crawler. Basically, Intel-Crawl is a tool that crawls the web to extract documents of interest using a previously developed library of semantic signatures. The library of semantic signatures is developed using the relevant training and testing data sets.

Each stage is explained in detail below.

Building a training set

For building the training set, a number of sample URLs related to a certain topic are collected and processed by the primary stage of Intel-Crawl, which extracts the content (text) that is present between the paragraph tags in the web page and links present in the webpage.

Building testing set

For building the testing set, a number of arbitrary seeds are selected and processed by the primary stage of Intel-Crawl, which will extract the content (text) and links present the web page.

Generation of semantic signature libraries

This module mainly uses Automated SSMinT to generate a set of semantic signatures using the training and testing sets. Automated SSMinT consists of four tools: Automated Keyword Tool (AKT), Automated Learner Tool (ALT), Hits Array Generator Tool (HAGT), and Semantic Signature Refinement Tool (SSRT).

Folders (Training and Testing), which were already built, are given as input to Automated Keyword Tool (AKT), which uses the TF-IDF technique; Term frequencies are computed in the training documents and the collection of testing documents is used to calculate inverse document frequencies. Keyword groups are constructed and stored in their respective Keyword Descriptor Files (KDFs). These KDFs are given as input to Automated Learner Tool (ALT), which generates the strong semantic signatures based on these keyword groups using the cosine distance measure. These semantic signatures will be stored in Semantic Signature Descriptor Files (SSDFs). Thus generated semantic signatures are not refined and they may not embody strong semantic signature content, so they are further processed in Hits Array Generator Tool (HAGT) where the Hits Array is generated using the semantic signatures as columns and training documents as rows The Hits Array is stored in a Hits Array Descriptor File (HADF). The Hits Array will be used for the refinement of semantic signatures using Semantic Signature Refinement Tool (SSRT).

Web crawler unit

A web crawler takes an input webpage and crawls through it to get the content of our interest. In the process of fetching the information required it will go through different phases. First, using the web link it will download the whole web page content. From this, it will fetch the text from the web page that is present between the paragraph tags. Then the processing of this text will be done by extracting all the links from it and removing all other HTML tags and other special characters used. Then this text is used to generate a row of the Hits Array in HAGT, where the semantic signatures are columns and the document that contains the extracted text is represented in the row. The row stores the semantic signature hits for the extracted text. If a nonzero row is obtained, then the link to this text is stored in a list that also contains the hits vector.

In our experiments, from the documents of links extracted, only the first ten links will be used initially, and for the each link the above described process is repeated and again links are extracted from each of the links and the first ten links are copied to a queue. Then the loop continues until our desired number of relevant web pages is met. This pruning of the web crawl was used to control the number of web pages logged and make the crawl finite. The number "ten" was chosen for convenience; different limits can be specified.

Working of the Web Crawler Unit

Intel-Crawl takes as input a web link and creates an HTTP connection. The content of the web page is downloaded the using an internet connection. Then text is extracted from the web page. Generally, text may be present in various formats such as Paragraph text, Tabular text, and Listed text. All these are extracted from the web page by the matching tags technique. Thus, the text extracted is used for our work. The text will be processed to refine the unwanted data such as HTML tags, special characters, and other HTML data which is used for styling of text. Before processing the text data, the links present in the text are extracted and stored separately for further crawling. In crawling the web pages, we implemented the Breadth First Search algorithm to visit the nodes (web pages) by following links. The flowchart of the entire web crawler unit of is shown in Figure 6.

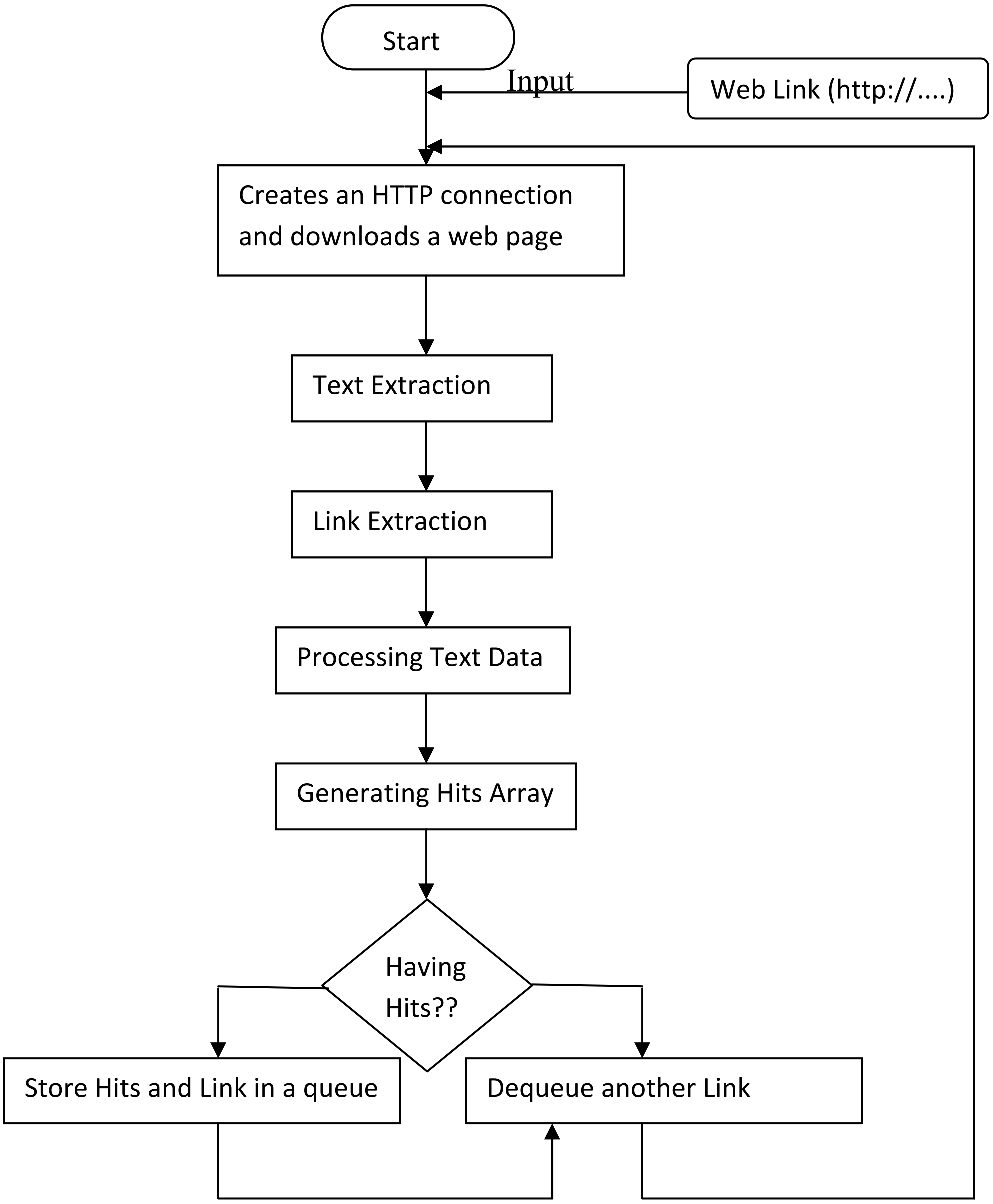

Figure 6. Flow of process in the web crawler unit

Creating a HTTP connection and downloading a web page

In this the HTTP link with different extensions like .html, .htm, .aspx, .asp, .php etc. are given as inputs, then it creates a HTTP connection with the server specified in the address and downloads the entire webpage using the TCP/IP protocol and the similar process will be repeated whenever a link was encountered and to be downloaded.

Text extraction

This module is the main challenging task where the text in a html web page can be of different formats since html gives the freedom of styling the html page to its user. Generally, the text may be present in between a paragraph tags, i.e. <p> </p>. If this is the case, then it can be easily extracted; however, there may be a lot of text written in different parts of the body. For example, some of the text may be written in <p> which does not have any end tags, some text may be between <P> and </P> where p is capitalized, and here also it may or may not have end tags. Some text may be in list tags <li> </li>, in order to mark it as listed data or in tabular format. So we need to extract all the useful text from all these tags in order to make the extracted text complete.

Here we are considering an example of a web page whose content is downloaded and the text extraction process is shown below.

```
</table>

<p> <b>Abortion</b> is defined as the termination of <a href="/wiki/Pregnancy_(mammals)"
title="Pregnancy (mammals)">pregnancy</a> by the removal or expulsion from the uterus of a <a
href="/wiki/Fetus" title="Fetus">fetus</a> or <a href="/wiki/Embryo" title="Embryo">embryo</a>
prior to <a href="/wiki/Fetal_viability" title="Fetal viability">viability</a>.<sup id="cite_ref-
definition_0-0" class="reference"><a href="#cite_note-definition-0"><span>[</span>note
1<span>]</span></a></sup> An abortion can occur spontaneously, in which case it is usually called
a <a href="/wiki/Miscarriage" title="Miscarriage">miscarriage</a>, or it can be purposely <a
href="//en.wiktionary.org/wiki/induce" class="extiw" title="wikt:induce">induced</a>. The term
<i>abortion</i> most commonly refers to the induced abortion of a human pregnancy.</p>
```

In the above example, is a part of HTML code for the page www.wikipedia.org/Abortion. In this module, the text present in the red is extracted since it is between the <p></p> tags.

Link extraction

All the links in a web page are extracted using a simple <href> tag as a reference. But, the main challenge occurs in the type of format in which the link is written. There are two types of referencing for a web page: link direct referencing and internal referencing. Direct referencing does not present any problem since it contains the whole link. All we need to do is extract it and copy it to a text document. Whereas, internal referencing has a problem. It does not specify the parent directory. Rather it will just specify the reference to the next page, assuming the current directory without including it explicitly. In this case, the parent directory has to be added every time in crawling the web pages.

```
</table>

<p> <b>Abortion</b> is defined as the termination of <a
href="/wiki/Pregnancy_(mammals)" title="Pregnancy (mammals)">pregnancy</a> by the
removal or expulsion from the uterus of a <a href="/wiki/Fetus" title="Fetus">fetus</a> or
<a href="/wiki/Embryo" title="Embryo">embryo</a> prior to <a href="/wiki/Fetal_viability"
title="Fetal viability">viability</a>.<sup id="cite_ref-definition_0-0" class="reference"><a
href="#cite_note-definition-0"><span>[</span>note 1<span>]</span></a></sup> An
abortion can occur spontaneously, in which case it is usually called a <a
href="/wiki/Miscarriage" title="Miscarriage">miscarriage</a>, or it can be purposely <a
href="//en.wiktionary.org/wiki/induce" class="extiw" title="wikt:induce">induced</a>. The
term <i>abortion</i> most commonly refers to the induced abortion of a human
pregnancy.</p>
```

In this module, the links present between the paragraph tags are extracted. They have the '*href*' tag. So all the relevant links in the HTML page are extracted.

Processing text data

This is a simple module that will remove all the unwanted tags and special characters present in the extracted text. The remaining plain text is stored in a document and the extracted text for each paragraph is appended to the text in the document.

```
</table>

<p> <b>Abortion</b> is defined as the termination of <a
href="/wiki/Pregnancy_(mammals)" title="Pregnancy (mammals)">pregnancy</a> by the
removal or expulsion from the uterus of a <a href="/wiki/Fetus" title="Fetus">fetus</a> or
<a href="/wiki/Embryo" title="Embryo">embryo</a> prior to <a href="/wiki/Fetal_viability"
title="Fetal viability">viability</a>.<sup id="cite_ref-definition_0-0" class="reference"><a
href="#cite_note-definition-0"><span>[</span>note 1<span>]</span></a></sup> An
abortion can occur spontaneously, in which case it is usually called a <a
href="/wiki/Miscarriage" title="Miscarriage">miscarriage</a>, or it can be purposely <a
href="//en.wiktionary.org/wiki/induce" class="extiw" title="wikt:induce">induced</a>. The
term <i>abortion</i> most commonly refers to the induced abortion of a human
pregnancy.</p>
```

```
Abortion is defined as the termination of pregnancy by the removal or expulsion from the
uterus of a fetus or embryo prior to viability. An abortion can occur spontaneously, in
which case it is usually called a miscarriage, or it can be purposely induced. The term
abortion most commonly refers to the induced abortion of a human pregnancy.
```

Generating the Hits Array

A matrix with columns corresponding to semantic signatures and rows corresponding to the documents from which text was extracted is called the Hits Array. For the generation of hits a cosine distance is measured between the centroid of the semantic signature and the vectors generated from the document. For all the vectors in a document, hits are counted and stored in a row of the Hits Array. The Hits Array is written in a Hits Array Descriptor File (HADF). The HADF is stored with .hadf extension and it is in the xml format which contains the link that was crawled and the set of semantic signatures that were used for calculating hits.

```
<HitsArrayDescriptorFile version="1.1">

<ssdSource folder="yes" file="no"></ssdSource>

<dataSource folder="no" file="yes"></dataSource>

<link>

http://en.wikipedia.org/wiki/Abortion_in_the_United_States

</link>

<ssds>

E:\Web Crawler\SSD library\tr_abortions_abortion_illegal_1 (2).ssd

E:\Web Crawler\SSD library\tr_abortions_abortion_illegal_1.ssd

E:\Web Crawler\SSD library\tr_abortions_rights_abortion_1.ssd

</ssds>

<HitsArray>

1 2 0 0 0 0 0 0 0 0 0 0 6 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 3
0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 6 9 0 1 7 2 0 0 0 0 0 2 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0

</HitsArray>
```

The example above shows the Hits Array Descriptor File structure which is in xml format. The link which was crawled is stored under the <link> tag and the Hits Array is stored under the <HitsArray> tag. Semantic signature libraries are stored under the <semantic signatures> tag.

Whenever the generated row of the Hits Array for a document has at least one nonzero entry, it is stored in the Hits Array and the link to the web page is placed in a queue. Otherwise, the web page that generated the text is ignored. The next link is de-queued from the queue and the process is repeated on the new link.

5. Experiments and Results

This chapter gives an overview of the experiments conducted using Intel-Crawl. All the experiments were performed based on the web links searched and the corpora was self-developed.

Experiment 1

Analysis of the performance of Intel-Crawl when keyword sets and semantic signatures generated by Automated SSMinT are used.
Case: keywords selected using TF-IDF.

Aim of the experiment

Evaluation of the performance of Intel-Crawl by crawling the input web links over the single topic ‘abortion’ using the keywords and semantic signatures generated by Automated SSMinT. Automated SSMinT selects keyword sets using TD-IDF, generates document vectors which store data on the frequencies and relative positions of the keywords in a given document, clusters these vectors to form a set of semantic signature candidates, refines the semantic signatures and prunes the set to create a library of semantic signatures that quantitatively capture the content of the training documents. The content of a web page is characterized by how frequently the various semantic signatures appear in the text derived from the web page.

Design of Experiment

For this experiment, a library of semantic signatures was developed based on the web links that were selected manually. The goal was to select web pages on a single topic, but include a spectrum of subgenres of the topic. Training set documents were extracted from fifteen web pages on the topic of abortion. For the testing set, which was used to calculate the IDF, twenty five URLs were selected manually based on random topics such as politics, sports, technology, health and science. The training and testing documents were used by Automated SSMinT to generate a library of semantic signatures. These semantic signatures were then used to characterize the content of newly visited web pages in Intel-Crawl.

First, 15 manually chosen web links were searched based on topic ‘abortion’, and then those links were passed to Automated SSMinT to generate our own library of semantic signatures. First, the links were given to web crawler to fetch the textual data and the web links that were in a web page and stored in a separate folders. Then the textual data was passed to Automated Keyword Tool (AKT). AKT generates keyword groups that are stored in separate Keyword Descriptor Files (KDFs). The generated KDFs are passed to Automated Learner Tool (ALT) to generate semantic signature descriptor files (SSDFs). These semantic signatures are not refined and based on all the groups of keywords, which may contain redundancy. So in order to refine them they were passed to Hits Array Generator (HAG), which produces a Hits Array Descriptor File (HADF). The HADF is an array of semantic signatures and training documents in columns and rows, respectively. The HADF is sent to Semantic Signature Refinement (SSR) for further refinement of semantic signatures. A library of forty-two refined semantic signatures, which were used as the basis for our experiments in the generation of the Hits Array.

In this experiment we chose the topic of interest as abortion. After a lot of manual research on the topic was conducted, we identified four major seeds in the topic of abortion: pro-life opinions and editorials, pro-choice opinions and editorials, abortion medicine, political legislation on abortion. Pro-choice opinion supports abortion rights, pro-choice opinion opposes abortion and often includes religion based thought, abortion medicine deals the medical procedures and terminology used in abortion, and political legislation on abortion concerns the laws and rights dealing with abortion. For each of the four topics a seed URL was chosen from the internet manually. URL for web pages that included good training text and a number of links to other related sites were chosen. For this experiment following URLs are used from each seed

Pro-Choice: http://www.whyprolife.com/pro-choice-arguments/

Pro-life: http://www.lifenews.com/

Abortion medicine: http://www.fwhc.org/abortion/index.htm

Political legislation: http://en.wikipedia.org/wiki/Abortion_in_the_United_States

Each URL is given as an input to the web crawler tool, so that it will fetch both the textual and other web links information present in a web page. First, it will fetch all the textual data present between <p> and </p> tags and <li> and </li> tags as most of the useful information is written between those two tags in a web page, all the web links that are used to reference to some other page will also be fetched and stored in a queue. The crawler will repeat the process for each of the links stored in the queue until six links, including the input seed URL, are followed to pages with the desired content. Here we restricted the crawl to six documents or links with semantic signature hits to facilitate the manual analysis on the results.

Results

Twenty-four Hit Array rows were generated and stored. Then the row vectors of the Hits Array were clustered using K-means clustering.

We used K-means clustering with K = 4, by generally thinking that we selected four seeds as our initial data. Cosine distance measure 'CD' is used in both the generation of semantic signature hits and in the clustering. The clustering results are shown in the following text box.

Cluster Number	Document Number	Description of the cluster
0	0,3,4,5,6,7,8,9,10,11,18,19,21	Pro-life and laws on abortion
1	12,14,16,20	Abortion medicine
2	13,15,17	Laws and rights about abortion
3	1,2,22,23	Laws, statistics and surveys about abortion

Table 1. Cluster analysis of web pages based on keywords and semantic signatures based on TF-IDF

The table below gives the URLs that were encountered by Intel-Crawl and also received semantic signature hits. The table gives a human analysis of the content of the web pages. The library of semantic signatures created is documented in the Appendix section of this document.

Doc No.	URL
	Keywords involved
0	http://www.whyprolife.com/pro-choice-arguments/
	Pro-life, pro-choice arguments, fetus, embryo, pregnancy, abortion, womb, force, legal abortion, Roe v Wade, Americans Predominant content: Pro-life, Pro-choice, laws, abortion medical terms
1	http://www.gallup.com/poll/20203/Americans-Favor-Parental-Involvement-Teen-Abortion-Decisions.aspx
	Parental involvement, teen abortions, law, abortion, supreme court, statistics and survey about abortion Predominant content: Statistics, surveys, laws about abortion
2	http://www.johnstonsarchive.net/policy/abortion/abreasons.html
	Abortion issue, law, debate, rape, US abortion legislation, legal abortion, statistics on abortion, rapes, abortion for sex selection Predominant content: Law about abortion, statistics on abortion
3	http://abortionviolence.com
	Pro – choice Vs. Pro – life, violence, pro – lifers, abortion Predominant content: Debate for Pro – Choice Vs. Pro – life

4	http://209.157.64.200/focus/f-news/1384733/posts
	Rudolph, bombing, Bible, church, Roman catholic Predominant content: Religious nature, Pro – life abortion
5	http://www.aboutabortions.com/Confess.html
	Abortions, abortion laws, supreme court, permissive abortion, catholic church, religions Predominant content: Religious nature, laws about abortion
6	http://www.lifenews.com/
	Planned parenthood, pro – life abortion Predominant content: pro – life abortion
7	http://www.lifenews.com/2012/07/05/obama-admin-sends-395k-to-tennessee-planned-parenthood/
	Pro – life, abortion, planned parenthood, US president Predominant content: pro – life abortion
8	http://www.lifenews.com/2012/07/05/obama-admin-sends-395k-to-tennessee-planned-parenthood/
	Pro – life, abortion, planned parenthood, US president Predominantly talks about: pro – life abortion
9	http://www.lifenews.com/2012/07/02/judge-blocks-mississippi-law-to-make-it-abortion-free/
	Law, abortion free, abortion, regulations, statistics of abortion, medical emergency Predominantly talks about: Laws about abortion
10	http://www.lifenews.com/2012/07/03/tennessee-new-pro-life-laws-on-abortion-take-effect/
	Pro – life laws, abortion, supreme court decisions, US, legalization Predominantly talks about: Pro – life abortion, laws about abortion
11	http://liveactionnews.org/opinion/abortion-men-you-have-a-say/
	Pregnant women, abortion, pro – choice, abortion debate, pro – lifers, pro – choicers Predominantly talks about: Pro – life Vs. Pro – choice abortions
12	http://www.fwhc.org/abortion/index.htm
	Abortion, legal, abortion pills, abortion information, post – abortion – syndrome, US researchers, US, abortion clinics, rights and laws of abortion, abortion statistics Predominantly talks about : medical terms in abortion
13	http://fwhc.org/espanol/index.htm
	This is the Spanish translated page of link in document number 12 which has same key words to describe
14	http://iamdrtiller.com/
	Safety, legal, abortion, abortion funds, abortion doctors, education about abortion Predominantly talks about: medical terms of abortion and education about abortion.
15	http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=US&vol=410&invol=113
	Roe.V.Wade, abortion laws, pregnancy, parenthood, law and court order

	regarding Roe.V.Wade Predominantly talks about: The law that court has passed regarding Roe.V.Wade
16	http://www.naral.org
	This is a news portal which gives information about pro – life, pro – choice and anti – abortion.
17	http://www.prochoiceactionnetwork-canada.org/articles/canada.shtml
	Laws of abortion, politicians, anti – choice harassment, pro – life, anti – abortion Predominantly talks about: laws of abortion
18	http://en.wikipedia.org/wiki/Abortion_in_the_United_States
	Supreme court, Roe. V. Wade, fetus, planned parenthood, fetal viability, women’s autonomy, legal abortion Predominantly talks about: Laws of abortion, pro – life Vs. pro – choice debate
19	http://en.wikipedia.org/wiki/Roe_v._Wade
	Roe. V. Wade, pro – life, pro – choice, laws of abortion, privacy Predominantly talks about: Pro- life Vs. Pro – choice abortion, laws of abortion
20	http://en.wikipedia.org/wiki/Planned_Parenthood_v._Casey
	Informed consent, abortion cases, court, Roe. V. Wade, planned parenthood Predominantly talks about: Informed consent and planned parenthood ofabortion
21	http://en.wikipedia.org/wiki/Roe_v._Wade
	Roe. V. Wade, pro – life, pro – choice, laws of abortion, privacy Predominantly talks about: Pro- life Vs. Pro – choice abortion, laws of abortion
22	http://en.wikipedia.org/wiki/Abortion
	Complete information about information that includes medical terms, laws of abortion, statistics about abortion Predominantly talks about: medical terms, Statistics of abortion
23	http://en.wikipedia.org/wiki/Common_law
	Laws in different countries, common laws Predominantly talks about: Laws in various countries

Table 2. URLs crawled by the web crawler using the seeds given as input

All the keywords involved in the above table are selected manually based on their importance to the document and their importance order was not sorted.

Interpretation of the results

In this experiment, Cluster 0 consists documents that have dominant content of pro-life, pro-choice, and abortion laws. Cluster 1 is grouped based on the pure medical terms and education regarding the abortion. Cluster 2 is grouped based on pure laws of abortion and other laws. Cluster 3 is formed mainly due to the statistical content present in the web pages and also based on the laws they discussed regarding abortion.

In this, document 23 has nothing to do with abortion, but it got hits because it is dominantly about laws and legalization of various common laws in various countries and. Though document 22 is purely concerned about abortion, it was included in Cluster 4 since it mainly contains statistics on abortion. In Cluster 0 all the pro-life and pro-choice op/ed pages are grouped, eventhough these subtopics are pointed in opposite directions, because the content of these web pages is neither opposing nor supporting one direction.

Experiment 2

Analysis of the performance of Intel-Crawl when keyword sets and semantic signatures generated by Automated SSMinT are used.
Case: keywords selected using term frequency alone.

Aim of the experiment

Evaluation of the performance of Intel-Crawl by crawling the web over the single topic ‘abortion’ using the keywords and semantic signatures generated by Automated SSMinT where term frequency alone is used for selecting keyword sets.

Design of experiment

In this experiment the database was developed by us by creating a library of semantic signatures using the keywords generated by Automated Keyword Tool (AKT) using the most frequent terms instead of Term Frequency – Inverse Document Frequency. Similar training and testing samples are used in the above experiment are used here.

In this experiment, using the fifteen training set documents that were retrieved on topic called abortion and the four hundred two testing set documents acquired on random topics were passed to Automated Keyword Tool (AKT) to generate a Key Word Descriptor File (KDF), but here instead of the previously used Term Frequency – Inverse Document Frequency, keywords are selected based on the most frequent terms that appear in the document. Thus selected keywords are grouped and written in to Keyword Descriptor File (KDF) and then these KDFs are parsed to Automated Learner Tool (ALT) to generate SSD (Semantic Signature Descriptor)s that are based on those keywords that are generated by most frequent terms and thus generated Semantic signatures are parsed along with the training document to obtain a Hits Array which comprises of a matrix that have documents as rows and Semantic signatures as columns and thus generated semantic signatures have redundant data that adds extra noise to the semantic signature generated and these are refined by parsing Semantic Signature Refinement Tool (SSRT) to refine these Semantic signatures to generate a final library of seventy Semantic signatures.

The same four seeds about abortion pro – choice, pro – life, medical abortion and political legislation of abortion are chosen in order to compare the current results with that of the previous experiment. These URLs are parsed to the Webcrawler to generate five documents that having hits to generate twenty four documents that having hits based on the generated library of Semantic signatures created by frequent terms as keywords.

Results

The documents generated are used to generate Hits array using the web crawler tool and those hits array descriptor files (HADFs) are merged into a single HADF and then parsed again to Semantic Signature Refinement Tool (SSRT) to generate a cluster using K-means clustering with K value as 4. The cluster results are tabulated in the following table.

Cluster No.	Document Nos.
0	0,3,4,5,6,7,8,9,10,11,16,18,19,21
1	12,20,22
2	13,15
3	1,2,14,17,23

Table 3. Cluster analysis of web links based on keywords and Semantic signatures developed based on TF-IDF

The URLs encountered and the keyword involved in the webpage and the topic that webpage is predominantly referring are tabulated in the following table and most of the URLs encountered are already on the earlier table, but for the comfort of the reader they are repeated again. The library of semantic signatures generated is documented in the Appendix section of this document.

Doc No.	URLs encountered
	Keywords Involved
0	http://www.whyprolife.com/pro-choice-arguments/
	Pro – life, pro – choice arguments, Fetus, embryo, pregnancy, abortion, womb, force, legal abortion, Roe. V. Wade, Americans Predominantly talks about : Pro life, Pro choice, laws, medical terms involved in abortion
1	http://www.gallup.com/poll/20203/Americans-Favor-Parental-Involvement-Teen-Abortion-Decisions.aspx
	Parental involvement, Teen abortions, law, abortion, supreme court, statistics and survey about abortion Predominantly talks about: Statistics, surveys, laws about abortion
2	http://www.johnstonsarchive.net/policy/abortion/abreasons.html
	Abortion issue, law, debate, rape, US abortion legislation, legal abortion, statistics on abortion, rapes, abortion for sex selection Predominantly talks about: Law about abortion, statistics on abortion
3	http://abortionviolence.com
	Pro – choice Vs. Pro – life, violence, pro – lifers, abortion Predominantly talks about: Debate for Pro – Choice Vs. Pro - life
4	http://209.157.64.200/focus/f-news/1384733/posts
	Rudolph, bombing, Bible, church, Roman catholic Predominantly talks about: Religious nature, Pro – life abortion
5	http://www.aboutabortions.com/Confess.html
	Abortions, abortion laws, supreme court, permissive abortion, catholic church,

	religions Predominantly talks about: Religious nature, laws about abortion
6	http://www.lifenews.com/
	Planned parenthood, pro – life abortion Predominantly talks about: pro – life abortion
7	http://www.lifenews.com/2012/07/05/obama-admin-sends-395k-to-tennessee-planned-parenthood/
	Pro – life, abortion, planned parenthood, US president Predominantly talks about: pro – life abortion
8	http://www.lifenews.com/2012/07/05/obama-admin-sends-395k-to-tennessee-planned-parenthood/
	Pro – life, abortion, planned parenthood, US president Predominantly talks about: pro – life abortion
9	http://www.lifenews.com/2012/07/05/european-parliament-adopts-resolution-condemning-forced-abortions/
	Parliament, forced abortions, sterilizations, abortions, pro – life Predominantly talking about Pro – life
10	http://www.lifenews.com/2012/07/02/judge-blocks-mississippi-law-to-make-it-abortion-free/
	Law, abortion free, abortion, regulations, statistics of abortion, medical emergency Predominantly talks about: Laws about abortion
11	http://www.lifenews.com/2012/07/03/tennessee-new-pro-life-laws-on-abortion-take-effect/
	Pro – life laws, abortion, supreme court decisions, US, legalization Predominantly talks about: Pro – life abortion, laws about abortion
12	http://www.fwhc.org/abortion/index.htm
	Abortion, legal, abortion pills, abortion information, post – abortion – syndrome, US researchers, US, abortion clinics, rights and laws of abortion, abortion statistics Predominantly talks about : medical terms in abortion
13	http://www.cedarriverclinics.org/
	Abortion care, birth control, women's health care, abortion, pregnancy test, birth control Predominantly talking about: abortion but less informative
14	http://fwhc.org/espanol/index.htm
	This is the Spanish translated page of link in document number 12 which has same key words to describe
15	http://www.cedarriverclinics.org/
	Abortion care, birth control, women's health care, abortion, pregnancy test, birth control Predominantly talking about: abortion but less informative
16	http://iamdrtiller.com/
	Safety, legal, abortion, abortion funds, abortion doctors, education about abortion Predominantly talks about: medical terms of abortion and education about

	abortion.
17	http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=US&vol=410&invol=113
	Roe.V.Wade, abortion laws, pregnancy, parenthood, law and court order regarding Roe.V.Wade Predominantly talks about: The law that court has passed regarding Roe.V.Wade
18	http://en.wikipedia.org/wiki/Abortion_in_the_United_States
	Supreme court, Roe. V. Wade, fetus, planned parenthood, fetal viability, women's autonomy, legal abortion Predominantly talks about: Laws of abortion, pro – life Vs. pro – choice debate
19	http://en.wikipedia.org/wiki/Roe_v._Wade
	Roe. V. Wade, pro – life, pro – choice, laws of abortion, privacy Predominantly talks about: Pro- life Vs. Pro – choice abortion, laws of abortion
20	http://en.wikipedia.org/wiki/Planned_Parenthood_v._Casey
	Informed consent, abortion cases, court, Roe. V. Wade, planned parenthood Predominantly talks about: Informed consent and planned parenthood of abortion
21	http://en.wikipedia.org/wiki/Roe_v._Wade
	Roe. V. Wade, pro – life, pro – choice, laws of abortion, privacy Predominantly talks about: Pro- life Vs. Pro – choice abortion, laws of abortion
22	http://en.wikipedia.org/wiki/Abortion
	Complete information about information that includes medical terms, laws of abortion, statistics about abortion Predominantly talks about: medical terms, Statistics of abortion
23	http://en.wikipedia.org/wiki/United_States_Constitution
	US, independence, laws, constitution Predominantly talks about: laws and constitution of US

Table 4. URLs crawled by the web crawler using the seeds given as input

Interpretation

Cluster 0 is mainly grouped based on the content of pro-life, pro-choice and laws on abortion. Cluster 1 is mainly grouped based on abortion medical terms. Cluster 2 is grouped based on similar documents that have content about abortion care. Cluster 3 is grouped based only on the laws of abortion and laws in general.

Here we chose frequently repeated words as keywords not based on their importance in the document, so there is a lot of irrelevance in getting hits, but it tried to cluster it in a good way. For example, documents 13 and 15 don't have much information, but a lot of repetitions of the word 'abortion', so hits were generated by the tool. These documents clustered into the same cluster. Remaining results are similar as those of the above experiment.

Experiment 3

Analysis of the performance of Intel-Crawl using a “bag of words” approach.
Case: keywords selected using TF-IDF.

Aim of the experiment

The aim of this experiment is to evaluate the performance of Intel-Crawl when a “bag of words” approach is used rather than the keyword sets and semantic signatures that are created by Automated SSMinT. A select bag of words was used; namely, the individual words of the keyword sets generated by Automated SSMint for the previous experiments. The content of a web page is characterized by how frequently these words appear in the text derived from the web page.

Design of the experiment

In this experiment a similar set of training and testing documents was used, where the training documents are based on abortion and testing corpora is based on five random seeds.

Using the testing and training corpora developed as the inputs to Automated Keyword Tool (AKT) keyword groups are generated and written in XML format in a file called Keyword Descriptor File (KDF) and thus written files are given as input to Automated Learner Tool (ALT) which gives us an output of Semantic Signatures written in a file called Semantic Signature Descriptor File (SSDFs) which are unfiltered ones that may have some unwanted Semantic signatures like redundant semantic signatures and noisy Semantic signatures and using those Semantic signatures as input to Hits Array Generator Tool (HAGT) hits array is generated using the Semantic signatures and Documents as columns and rows respectively and Thus generated hits array is stored in Hits Array Descriptor File (HADF) which are sent to Semantic Signature Refinement Tool (SSRT) to refine thus generated hits array. Then the refined forty two refined semantic signatures are used a library of semantic signatures in further crawling in using the Web crawler tool. Each of the SSDF is in XML format that having a group of keywords and the centroid and vector values, now in this experiment we are splitting those group of keywords into single words and the redundant words are removed to form a list of words that are used in generation of hits array. In this, a document is said to have a hit only if the content in that webpage is having at least one word in the bag of words. Here all the centroid techniques that were used earlier are ignored and just the comparison techniques are used in the generation of hits array. Here in our experiment we have a bag of fifty four non redundant keywords.

For the experiment of crawling we are using the same seed web links used in the first experiment in order to make this experiment compatible for the comparison between all the experiments.

Results

The hits array thus generated with the list of words as columns and documents as rows are stored in a hits array descriptor file (HADF) and now they are merged into a single hits array for the clustering using SSRT using the K-means clustering with the value of K as four. Thus generated clustering results are tabulated in the following table

Cluster No.	Document Nos.
0	9
1	0,1,2,3,4,5,6,7,8,10,11,12,13,16,17,18,19,20,21
2	14,15
3	22,23

Table 5. Clustering based on the bag of words using the keywords generated based on TF-IDF

The URLs that got hits based on the bag of words and the keywords present in that textual content are tabulated in the following table. All the bag of words used for this experiment are documented in the appendix section of this document.

Doc No.	URL encountered
	Keywords involved
0	http://www.whyprolife.com/pro-choice-arguments/
	Pro – life, pro – choice arguments, Fetus, embryo, pregnancy, abortion, womb, force, legal abortion, Roe. V. Wade, Americans Predominantly talks about : Pro life, Pro choice, laws, medical terms involved in abortion
1	http://www.gallup.com/poll/20203/Americans-Favor-Parental-Involvement-Teen-Abortion-Decisions.aspx
	Parental involvement, Teen abortions, law, abortion, supreme court, statistics and survey about abortion Predominantly talks about: Statistics, surveys, laws about abortion
2	http://www.johnstonsarchive.net/policy/abortion/abreasons.html
	Abortion issue, law, debate, rape, US abortion legislation, legal abortion, statistics on abortion, rapes, abortion for sex selection Predominantly talks about: Law about abortion, statistics on abortion
3	http://abortionviolence.com
	Pro – choice Vs. Pro – life, violence, pro – lifers, abortion Predominantly talks about: Debate for Pro – Choice Vs. Pro - life
4	http://www.nationalreview.com/comment/graham200403100905.asp
	Anti – abortion violence, abortion clinics, pro – life, planned parenthood, pro – choice, legal abortion, Roe. V. Wade, anti – abortion Predominantly talks about: Pro choice Vs. Pro – life abortion
5	http://209.157.64.200/focus/f-news/1384733/posts
	Rudolph, bombing, Bible, church, Roman catholic Predominantly talks about: Religious nature, Pro – life abortion
6	http://www.lifenews.com/
	Planned parenthood, pro – life abortion

	Predominantly talks about: pro – life abortion
7	http://www.lifenews.com/2012/07/05/obama-admin-sends-395k-to-tennessee-planned-parenthood/
	Pro – life, abortion, planned parenthood, US president Predominantly talks about: pro – life abortion
8	http://www.lifenews.com/2012/07/05/obama-admin-sends-395k-to-tennessee-planned-parenthood/
	Pro – life, abortion, planned parenthood, US president Predominantly talks about: pro – life abortion
9	http://www.lifenews.com/2012/07/05/catholics-for-choice-mocks-usccb-fortnight-for-freedom/
	Catholics, defense, abortion, religions rights of citizens Predominantly talks about: Religious nature and rights of citizens
10	http://www.lifenews.com/2012/07/05/european-parliament-adopts-resolution-condemning-forced-abortions/
	Parliament, forced abortions, sterilizations, abortions, pro – life Predominantly talking about Pro – life
11	http://www.lifenews.com/2012/07/02/judge-blocks-mississippi-law-to-make-it-abortion-free/
	Law, abortion free, abortion, regulations, statistics of abortion, medical emergency Predominantly talks about: Laws about abortion
12	http://www.fwhc.org/abortion/index.htm
	Abortion, legal, abortion pills, abortion information, post – abortion – syndrome, US researchers, US, abortion clinics, rights and laws of abortion, abortion statistics Predominantly talks about : medical terms in abortion
13	http://www.cedarriverclinics.org/
	Abortion care, birth control, women's health care, abortion, pregnancy test, birth control Predominantly talking about: abortion but less informative
14	http://fwhc.org/espanol/index.htm
	This is the Spanish translated page of link in document number 12 which has same key words to describe
15	http://www.glamour.com/sex-love-life/2009/02/eight-women-share-their-abortion-stories
	Pregnancy test, abortion, personal story Predominantly talks about: Personal story of different women how their abortion was done
16	http://www.rhrealitycheck.org/fact-v-fiction/abortion-causes-a-variety-of-health-complications
	Abortion, health risk, complications, breast cancer, depression, infertility, safe Predominantly talks about: Personal issues about abortion
17	http://www.cedarriverclinics.org/
	Abortion care, birth control, women's health care, abortion, pregnancy test,

	birth control Predominantly talking about: abortion but less informative
18	http://en.wikipedia.org/wiki/Abortion_in_the_United_States
	Supreme court, Roe. V. Wade, fetus, planned parenthood, fetal viability, women's autonomy, legal abortion Predominantly talks about: Laws of abortion, pro – life Vs. pro – choice debate
19	http://en.wikipedia.org/wiki/U.S._state
	Federal state, US, about US Predominantly talks about: US
20	http://en.wikipedia.org/wiki/Roe_v._Wade
	Roe. V. Wade, pro – life, pro – choice, laws of abortions, privacy Predominantly talks about: Pro – life Vs. Pro – choice abortion
21	http://en.wikipedia.org/wiki/Planned_Parenthood_v._Casey
	Informed consent, abortion cases, court, Roe. V. Wade, planned parenthood Predominantly talks about: Informed consent and planned parenthood of abortion
22	http://en.wikipedia.org/wiki/Strict_scrutiny
	US courts, strict scrutiny, supreme court, laws Predominantly talks about: laws of US
23	http://en.wikipedia.org/wiki/Strict_scrutiny
	US courts, strict scrutiny, supreme court, laws Predominantly talks about: laws of US

Table 6. URLs crawled by the web crawler using the seeds given as input.

The keywords stated above are decided manually by analyzing each and every link and the order of importance is not defined for any of the keyword sets.

Interpretation

In this experiment we chose individual words instead of groups of keywords. The clustering performance degrades since just a match of word in the web content results in a hit, but the importance of word in the whole document is not considered. In this, document 9 pops out in a separate cluster as it matches with the religious nature. Cluster 1 is grouped based on the concepts of pro-life and pro-choice. Cluster 2 groups the treatment, abortion, and medical terms of abortion. Cluster 3 contains the documents regarding laws in the US.

In this experiment, many of the pages are not directly related to the content of interest, though they have some point related to abortion. For example, documents 14 and 15 are the personal stories of various women that got an abortion for different reasons and how they obtained an abortion, which is subtopic content not related to the content of the training documents, which we chose. But documents 14 and 15 just have a word or two that match to generate a hit and so the web crawler retains them in its list.

Experiment 4

Analysis of the performance of Intel-Crawl using a “bag of words” approach.
Case: keywords selected using term frequency alone.

Aim of the experiment

This experiment is aimed to analyze how well the web crawler can show its performance in getting showing the results that are related to the topic of interest using the bag of words approach and the keyword group generated using the most frequent words in a document.

Design of experiment

This experiment uses the same database that the previous experiment uses, the fifteen training documents related to a single topic called abortion though they are varied by their sub topic and the test set is a random topic documents that are generated using twenty five documents, 5 from each of the random topics like science, health, politics, technology and sports. From them we developed a library of seventy semantic signatures.

This experiment was also conducted in a similar fashion as the earlier experiments by generating a set of Keyword Descriptor Files using the Automated Keyword Tool using the generated training and testing set and then those KDFs underwent to Automated learner tool to generate crude clusters of Semantic signatures these are parsed to Hits array generator tool to obtain Hits array descriptor which in turn passed to Semantic signature refinement tool to obtain refined seventy Semantic signatures. From each of the Semantic signatures words are separated and the redundant words are filtered and saved in list and pass to Web crawler along with the URLs of four seeds to obtain the web pages that have the content of our interest. That list of URLs along with their hits array is stored in a Hits array descriptor file and that is parsed to SSRT to obtain different clusters of data using the K-means algorithm with possible k value as 4.

Results

After parsing each of the web links stated in the first experiment totally 24 corresponding web links along with the data and hits array is generated and saved in a cluster and that is tabulated in the following table.

Cluster No.	Document No.
0	5,9,19
1	0,1,2,3,4,6,7,8,9,10,11,13,16,17,18,20,21
2	12,14,15,17
3	22,23

Table 7. Clustering based on the bag of words using the keywords generated based on TF-IDF

URLs used and keywords involved in each URL are specified in a detailed manner in the following table

Doc No.	URLs encountered
	Keywords involved
0	http://www.whyprolife.com/pro-choice-arguments/
	Pro – life, pro – choice arguments, Fetus, embryo, pregnancy, abortion, womb, force, legal abortion, Roe. V. Wade, Americans Predominantly talks about : Pro life, Pro choice, laws, medical terms involved in abortion
1	http://www.gallup.com/poll/20203/Americans-Favor-Parental-Involvement-Teen-Abortion-Decisions.aspx
	Parental involvement, Teen abortions, law, abortion, supreme court, statistics and survey about abortion Predominantly talks about: Statistics, surveys, laws about abortion
2	http://www.johnstonsarchive.net/policy/abortion/abreasons.html
	Abortion issue, law, debate, rape, US abortion legislation, legal abortion, statistics on abortion, rapes, abortion for sex selection Predominantly talks about: Law about abortion, statistics on abortion
3	http://abortionviolence.com
	Pro – choice Vs. Pro – life, violence, pro – lifers, abortion Predominantly talks about: Debate for Pro – Choice Vs. Pro - life
4	http://www.nationalreview.com/comment/graham200403100905.asp
	Anti – abortion violence, abortion clinics, pro – life, planned parenthood, pro – choice, legal abortion, Roe. V. Wade, anti – abortion Predominantly talks about: Pro choice Vs. Pro – life abortion
5	http://www.acf.hhs.gov/programs/cb/stats_research/afcars/tar/report14.htm
	Children's adoption, statistics and parental rights about abortion Predominantly talked about statistical and medical
6	http://www.lifenews.com/
	Planned parenthood, pro – life abortion Predominantly talks about: pro – life abortion
7	http://www.lifenews.com/2012/07/05/obama-admin-sends-395k-to-tennessee-planned-parenthood/
	Pro – life, abortion, planned parenthood, US president Predominantly talks about: pro – life abortion
8	http://www.lifenews.com/2012/07/05/obama-admin-sends-395k-to-tennessee-planned-parenthood/
	Pro – life, abortion, planned parenthood, US president Predominantly talks about: pro – life abortion
9	http://www.lifenews.com/2012/07/05/catholics-for-choice-mocks-usccb-fortnight-for-freedom/
	Catholics, defense, abortion, religions rights of citizens Predominantly talks about: Religious nature and rights of citizens
10	http://www.lifenews.com/2012/07/05/european-parliament-adopts-resolution-condemning-forced-abortions/

	Parliament, forced abortions, sterilizations, abortions, pro – life Predominantly talking about Pro – life
11	http://www.lifenews.com/2012/07/02/judge-blocks-mississippi-law-to-make-it-abortion-free/
	Law, abortion free, abortion, regulations, statistics of abortion, medical emergency Predominantly talks about: Laws about abortion
12	http://www.fwhc.org/abortion/index.htm
	Abortion, legal, abortion pills, abortion information, post – abortion – syndrome, US researchers, US, abortion clinics, rights and laws of abortion, abortion statistics Predominantly talks about : medical terms in abortion
13	http://www.cedarriverclinics.org/
	Abortion care, birth control, women's health care, abortion, pregnancy test, birth control Predominantly talking about: abortion but less informative
14	http://fwhc.org/espanol/index.htm
	This is the Spanish translated page of link in document number 12 which has same key words to describe
15	http://www.glamour.com/sex-love-life/2009/02/eight-women-share-their-abortion-stories
	Pregnancy test, abortion, personal story Predominantly talks about: Personal story of different women how their abortion was done
16	http://www.cedarriverclinics.org/
	Abortion care, birth control, women's health care, abortion, pregnancy test, birth control Predominantly talking about: abortion but less informative
17	http://iamdrtiller.com/
	Safety, legal, abortion, abortion funds, abortion doctors, education about abortion Predominantly talks about: medical terms of abortion and education about abortion.
18	http://en.wikipedia.org/wiki/Abortion_in_the_United_States
	Supreme court, Roe. V. Wade, fetus, planned parenthood, fetal viability, women's autonomy, legal abortion Predominantly talks about: Laws of abortion, pro – life Vs. pro – choice debate
19	http://en.wikipedia.org/wiki/U.S._state
	Federal state, US, about US Predominantly talks about: US
20	http://en.wikipedia.org/wiki/Roe_v._Wade
	Roe. V. Wade, pro – life, pro – choice, laws of abortions, privacy Predominantly talks about: Pro – life Vs. Pro – choice abortion
21	http://en.wikipedia.org/wiki/Planned_Parenthood_v._Casey
	Informed consent, abortion cases, court, Roe. V. Wade, planned parenthood

	Predominantly talks about: Informed consent and planned parenthood of abortion
22	http://en.wikipedia.org/wiki/Strict_scrutiny
	US courts, strict scrutiny, supreme court, laws Predominantly talks about: laws of US
23	http://en.wikipedia.org/wiki/Strict_scrutiny
	US courts, strict scrutiny, supreme court, laws Predominantly talks about: laws of US

Table 8. URLs crawled by the web crawler using the seeds given as input

All the words are manually chosen by visiting each and every link is manually analyzed to find the keywords and the order of keywords are not specified.

Interpretation

This experiment results in the least clustering possible as both the group of words are violated and importance of words in a document is also violated results in the above table of clustering where the documents in each clusters have least correlation to be in a single cluster. Almost all the documents which got hits are not completely related to a single word.

Summary analysis of the experiments

Based on all the experiments conducted, we drew the following conclusions

- Generation of semantic signatures and keywords using TF-IDF is the best clustering approach since all the documents that got hits are properly related to the content of interest. For example in Experiment 1, document 5 is not directly related to pro-life but the words in it are strongly related to pro-life.
- Generation of semantic signatures and keywords using most the frequent terms approach gave a similar clustering result when compared with Experiment. However, document 13 and 15 are less informative and they are completely related to customer care information regarding abortion. Since they have the word ‘abortion’ repeated more times these web links were considered contain the content of interest. Similarly, document 11 is also less related to pro-life, but due to the fact that the term catholic is repeated many times, hits were generated and it was clustered in the pro-life group, which is not an accurate grouping.
- In the “bag of words” approach with keywords generated using TF-IDF, many of the clustered documents are irrelevant. If we consider 14 and 15 for example, both are just the personal stories for few women about how they obtained an abortion and how they use the contraceptive methods to avoid pregnancy, etc., since the list of words have exact match on the web pages. But, this subtopic area was not targeted by the user who chose the training documents.

- In the final experiment, the “bag of words” approach with keywords generated using most frequent words, almost all the clusters are less meaningful than in the other experiments as they groups were formed based on the matching of a single word without using any relational data between words.

The major observation we made from these experiments is that as the web crawler crawled in every experiment, the experiments using the “bag of words” approach returned documents with less discrimination and content accuracy than for the experiments using semantic signatures to characterize document content. The broadly range of documents returned added noise in clustering phase of data analysis, which caused ill-defined clusters. The semantic signature driven intelligent web crawling was far superior.

6. Conclusions and Future Work

We were successful in developing a web crawler that crawls over the web to retrieve the information related to the user’s interest. Before conducting experiments with the web crawler we developed a library of semantic signatures using the keyword groups generated using two sets of documents, first set is related to the topic of interest and second set is the random documents from the corpus (internet). All the information regarding those set of documents are clearly given in the appendix of this document.

Intel-Crawl uses the semantic signatures developed and crawls over the web to acquire the data by visiting each web page and check whether a ‘hit’ occurs in that page or not. If there is a hit, the corresponding value in the Hits Array will be incremented. All the web pages are initially saved in a breadth first search tree and scrutinized later and all the web pages that have hits will be stored in a separate list of URLs along with the hits.

An advantage of our crawler is that it can extract text information from almost all kinds of web page formats and can be used in monitoring a web page irrespective of the type of content and whether the content is constant or dynamic. Here the functionality of Automated SSMinT is extended to web data, and the semantic signatures produced by Automated SSMinT are employed to intelligently crawl the web. This semantically informed web crawling is successful in identifying web pages with targeted content.

We conducted experiments that changed various parameters and different keyword generation techniques to observe the effectiveness of the proposed technique over other techniques that can generate similar kind of results. At last, we can clearly observe that the content characterization of data using semantic signatures and generating the keyword groups using the Term Frequency – Inverse Document Frequency weight technique has better results than using the “bag of words” approach and other keyword group generation techniques such as most frequent words.

We would like to extend the functionality and the application of this web crawler in various scenarios.

- Use the intelligent web crawler to crawl over various blogs to identify violent and illegal intent.
- By acquiring information over a period of time from a website we can identify changes and signature patterns towards a single strong concept to conclude its organization's activity.
- Enhanced Intel-Crawl to also generate hyperlink networks in which web pages are nodes and links are retained (or assigned a weight attribute) based on the content of the linked document. Thus, World Wide Web content specific link networks can be built and further studied using social network methodologies. Crawls from targeted content seed websites repeated over time can be used to monitor shifts in web page content and link network evolution.

Appendix

List of Semantic Signature Descriptor files used:

tr_abortion_biblical_articles_1

tr_abortion_cases_issue_1

tr_abortion_cases_some_1

tr_abortion_choice_pro_1

tr_abortion_choice_pro_2

tr_abortion_definition_history_1

tr_abortion_issue_some_1

tr_abortion_laws_legal_1

tr_abortion_rights_pro_1

tr_abortion_violence_anti_1

tr_abortion_women_pregnancy_1

tr_abortions_abortion_illegal_1 (2)

tr_abortions_legal_abortion_1

tr_abortions_women_obtain_2

tr_christ_death_corinthians_1

tr_christ_jesus_death_1

tr_fetus_choice_life_1

tr_god_david_life_1

tr_god_people_justice_1

tr_human_cell_single_1

tr_life_quality_child_1

tr_life_some_people_2

tr_parents_people_children_2

tr_person_human_rights_1

tr_person_human_zygote_1

tr_potential_dna_single_1

tr_potential_human_dna_1

tr_potential_human_zygote_1

tr_potential_life_body_1

tr_pregnancy_women_states_1

tr_pro_abortion_choice_1

tr_pro_abortion_life_1 (2)

tr_pro_just_choice_1

tr_pro_legal_life_1

tr_pro_life_just_1

tr_psalm_womb_mother_1

tr_questions_abortion_asked_1

tr_right_person_life_1

tr_rights_abortion_states_1

tr_teaches_bible_human_1

tr_teaches_life_human_1

tr_teaches_womb_child_1

tr_teaches_womb_human_1

tr_woman_fetus_responsible_1

tr_women_abortion_states_1

tr_women_feminists_life_1

tr_zygote_cell_single_1

tr_zygote_human_dna_1

Bibliography

[1] Wikipedia. (2010), Text mining --- Wikipedia, the free encyclopedia. Available: http://en.wikipedia.org/wiki/Text_mining

[2] Ravali Kota. Automated Discovery of Relevant Features for Text Mining. Master's Thesis, Lane Dept. Computer Science and Eng., West Virginia University. 2010.

[3] Uday Kiran Para. Computer-aided semantic signature identification and document classification via semantic signatures. Master's Thesis, Lane Dept. Computer Science and Eng., West Virginia University., 2010.

[4] Sri Ramya Peddada. Sensitivity of semantic signatures in text mining. Master's Thesis, Lane Dept. Computer Science and Eng., West Virginia University. 2010.

[5] Wikipedia (2010), Web Crawler --- Wikipedia, the free encyclopedia. Available: http://en.wikipedia.org/wiki/Web_crawler

[6] Wikipedia (2011), tf-idf --- Wikipedia, the free encyclopedia. Available: http://en.wikipedia.org/wiki/Tf-idf

[7] http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

[8] Wikipedia. K-means clustering. http://en.wikipedia.org/wiki/K-means_clustering

[9] K-Means Clustering Algorithm. http://cs.gmu.edu/cne/modules/dau/stat/clustgalgs/clust5_bdy.html

[10] S. Weiss, N. Indurkhya, T. Zhang and F. Damerau. (2004, October). Text Mining: Predictive Methods for Analyzing Unstructured Information Available: http://www.worldcat.org/isbn/0387954333

[11] Wikipedia. (2010), Cosine Similarity --- Wikipedia, the free encyclopedia. Available: http://en.wikipedia.org/wiki/Cosine_similarity

[12] Wikipedia. (2010), Singular value decompositon --- Wikipedia, the free encyclopedia. Available: http://en.wikipedia.org/wiki/Singular_value_decomposition

[13] Princeton University, "Breadth-first search?"

[14] Hai Dong , Farookh Khadir Hussain: "Focused Crawling for Automatics Service Discovery, Annotation, and Classification in Industrial Digital EcoSystems" , *IEEE Transactions*, Vol 58, No.6, June 2011.

[15] E. Francesconi and G. Peruginelli: "Searching and retrieving legal literature through automated semantic indexing," *Proc. ICAIL*, 2007, pp. 131–138.

[16] C. L. Giles, Y. Petinot, P. B. Teregowda, H. Han, S. Lawrence, A. Rangaswamy, and N. Pal, "eBizSearch: A niche search engine for e-business," *Proc. SIGIR*, 2003, pp. 413–414.