Development of a tool for the identification of counterfeiters on the internet
Raül Rabionet Janssen
Master Project Defense
Supervisor: Martin Rajman

The IP Project

- The Internet Patrolling Project is a collaboration between EPFL and VenTek International
- It started in June 2007 and is expected to finish in November 2008
- The main goal of the project is to develop a software that will be able to:
  - Detect web sites that deal with counterfeited products
  - Extract information that can lead to identifying the individuals behind this web sites
  - Classify the processed websites in clusters in order to unify all information on specific individuals
  - Fill in a template with all the useful information oriented to initiating the necessary legal actions
The Master Project

• This Master Project has been conceived to match the starting stages of the IP Project
• The Main work on these stages has been.
  – Design and specification of the solution
  – Preparation of a work environment (server, website, wiki)
  – Implementation of a full working first version of the prototype, called “Demonstrator”
  – Documentation of these steps

The FAST Platform

• The IP Project is being developed in collaboration with VenTek, a company specialized in document search
• Part of their contribution to the project is making available a powerful platform for document processing
• The current prototype has been strongly based on FAST
• Therefore, a part of the work involved in this Master Project has been understanding the platform and learning how to develop and debug code specifically intended for it.
Motivation

- Internet has become the greatest market for counterfeited and fake products
- Losses to worldwide companies are estimated close to $600 billion
- Legal actions against counterfeiters are hindered by international law issues and lack of evidence directing to individuals
- Although specific software products have been developed for internet monitoring, no specific details are provided on the extent of these tools

Implementation of the Demonstrator

- We have divided the design in 3 modules:
  - Module 1: Gathering and processing of documents. Starting with a set of keywords, we should obtain an index of processed documents and extracted entities.
  - Module 2: Link Based analysis and document clustering. Using the Entities to define the documents, we should be able to cluster together documents referring to the same counterfeiting agents.
  - Module 3: Information output. The information contained in each cluster should be output in a readable format.
Module 1: Web Search

- A set of keywords is used to start the Web Search, using:
  - Words identifying the affected company or products
  - Keywords related to counterfeit and gray market
- A common web search engine (Google) is used to retrieve pages that contain the chosen keywords
- After processing the retrieved pages, the most relevant Entities are fed again to the search engine, and further documents are processed
  - This Iterative Crawl produces documents not necessarily linked to counterfeiting
  - These documents could provide more accurate information to identify the individuals behind the counterfeiting activity through their presence in other areas on the web.
Module 1: Obtaining URLs

- A Java package (SeedGenerator) has been created by EPFL.
- It gathers the first n Google results for a set of given keywords.
- Receives the links from Google in HTML format, looking for the tag that identifies the results
  - An API used to exist to avoid this process, but it is currently only available to be embedded in websites.
- It provides a list of relevant URLs

---

Module 1: Gathering and processing documents

- The FAST platform is being used to crawl the web and process the documents.
- The crawl starts with the list of URLs provided by the SeedGenerator. The depth of the crawl will vary depending on which iteration is being executed.
  - The first crawl is being run 1 level deep
  - The iterative crawl is only aiming at the URLs identified by the Seed Generator
- The documents’ processing consists mainly in:
  - Adding the whois information of the website
  - Extracting the Entities:
    - Person and company names
    - Phone numbers
    - IP address & DNS name
    - E-mail addresses
  - Indexing the documents
Module 1: Scoring Relevant Pages

• We are assuming that counterfeiting activity will not be carried out in websites with a high popularity in the Internet.
• Therefore, a page is considered relevant if its popularity in the Internet is not too high.
  – DNS_Freq = #Hit on Google search.
  – Score = 1 / DNS_Freq
  – Score > 10^{-6} => relevant page.

This score can be useful for pages retrieved in the first iteration. In the second iteration we will be retrieving pages from very diverse sources and we might need a different scoring.

Module 1: Entities for next Iteration

• We want to choose Entities that:
  – Can lead to pages with more information on the counterfeiters or their activity (we want to choose Entities strongly linked to counterfeiting)
  – Will not produce a lot of noise
• We make the assumption that an Entity is potentially interesting if:
  – It appears in a page that was considered “relevant” according to the previous filter
  – Is not very frequent in the Internet in general
• We set a list of thresholds for each Entity, defined by experimentation. We feed for the Iteration Entities that:
  – Appear in a relevant page
  – Have a frequency on the Internet below the threshold
Module 1: Document Output

- We want to have each module work separately, to allow independent development and testing.
- After each document is processed in Module 1, its main elements are output in .XML format.
- These .xml files can be fed to a new FAST collection when module 2 is started.

Module 2

FAST Platform – Module 2

Collection of processed documents from Step 1 in XML format → Generation of docvectors → Selection of Entities relevant for clustering → SQL Scored Entity Database

Link – Based Analysis Pipeline

Import XML documents into FAST format → Insert Entity information and scoring in the docvector fields → Automatic docvector clustering → FAST Index

Labeling of each document with its cluster name → Cluster identification

FAST Index with cluster Labels
Module 2: Clustering

• FAST includes a clustering tool that uses the following process:
  – A document is summarized as a docvector, choosing pairs of words from its content.
  – The documents are then clustered based on similarity between the docvectors.

• We are defining new docvectors for each document, based on the Entities it contains.

• We are then using the same clustering process, based on similarity between Entity Vectors.

Module 2: Docvectors

• The docvectors that summarize a document have a structure such as:
  [International Politics, 0.87][Global Warming, 0.54][United Nations, 0.47]
  Each keyword is weighted according to its importance to the page.

• We define the same structure, using the Entities as keywords.

• The weights are computed as a reverse function of the frequency of the Entity on the Internet.

• Other weighting criteria will be considered in the future.
Module 2: Automatic Clustering

- FAST clustering uses an algorithm based on vector similarity:
  - Each cluster is represented by one of the documents in the cluster
  - A document belongs to the cluster to whose representative it has the highest similarity value, as long as this similarity is above a customizable threshold.
  - Most documents end up in a singleton cluster
- All documents are clustered using this algorithm. All documents forming a cluster receive a label with their cluster name.
- A navigator is defined in FAST’s Search View, in order to allow searching the clusters separately, and viewing all Entities that belong to a cluster.

Module 3: Multidocument Information Extraction

- The goal of the tool is to fill in a frame with all the information that can be useful for the prosecution of the counterfeiters
- This process of Information Extraction is currently using existing FAST functionalities, allowing navigation across the identified clusters
Tests: Test data

- We have worked on the development through a series of successive tests and improvements
- We have built a test data set that will be used during the rest of the development:
  - The main element is a list of initial websites, obtained by choosing pages similar to an initial set of accepted counterfeiting pages.
  - The initial pages have been manually classified as either “good” or “bad” based on their relevancy to the project. We need to keep the “bad” pages in the data set to test the ability of the demonstrator to filter out noise.
  - The initial set contains roughly 200 pages, from which around 50% have been classified as “good”.

Tests: Performance. Module 1

- The results of the last test we have run can be summarized as follows:
  - We started with 50 pages from the data set – we chose only “good” pages, in order to test the capabilities of the clustering tool provided the least noise possible.
  - We crawled from those pages, gathering 6143 documents
  - We selected 1083 Entities to be used for Iterative Crawl
  - We ended up with 19856 documents after Module 1
  - This process took about 24 hours
Tests: Performance. Module 2

- The results of the Module 1 where entered into Module 2 as .xml documents:
  - We identified 3 clusters, with
    - 3 documents
    - 7 documents
    - 3 documents
  - All of them where formed around an IP number and some other Entity
  - One of the clusters identified the following:
    - 209.34.176.11 to be the IP address of two separate counterfeiting websites (www.watchec.com and www.replica-watch-shopping.com)
    - Michelle Chen to be related to both web pages through their whois information
  - We believe that the way the docvectors were constructed was too restrictive (with too high weights and too many Entities), thus making the threshold of similarity too high.
  - By working on the weights of the Entities, better results should be achievable and are already being worked upon.

Tests: Identified Cluster

Raul Rabionet
Master Project Defense
Future Work: Module 1

- Entity Extraction: Errors discovered in the extraction process during the testing stage are being reported and will be corrected
  - Interesting Entities identified in pages that were not properly extracted
  - Extractions that do not correspond to real Entities.
- Entity Selection: The number of Entities currently selected for the Iteration is probably too large. We want to introduce some other criteria for selecting an Entity (such as its frequency inside the data set) to reduce the amount of noise.
- Entity Scoring: We want to define other scoring functions, probably based on the structure of the page, to improve the usefulness of the weights on the Entities.

Future Work: Module 2

- In the initial description of the project, Module 2 was intended to be implemented using Link-Based Analysis
- A collaboration has started with a team Hebrew University in Jerusalem, specialized in this technology
Future Work: Module 3

- Currently the extraction of the final results is being done through FASTs Search View
- Once the Case Frames have been properly described, it is intended that a specific tool will be developed.
- This tool will aid the user of the prototype in the selection of the proper output for each cluster