Name Disambiguation

Rich Ernst CSE 8331

Overview

- What is it?
- Relevant papers
  - A Brief Survey of Automatic Methods for Author Name Disambiguation (Ferreira et al)
  - Name Disambiguation in Author Citations using a K-way Spectral Clustering Method (Han et al) 2005
  - Fast Author Name Disambiguation in CiteSeer (Huang et al) 2006
  - Efficient Topic-based Unsupervised Name Disambiguation (Song et al) 2007
- Practical application for paper

What is name disambiguation?

- Initialized first name
  - J. Smith = John Smith, James Smith, John B. Smith etc
- AKA Names
  - Bill Jones = William Jones, Wilhelm Jones
- Same Name
  - Richard Ernst (InfoSec) = Richard Ernst (Nobel Chemist)
- Mispellings
  - Ian McTavish = Ian Mactavish, Ian McTavich
- Changes
  - Lillian Davis = Lillian Taylor (name change)

Papers

A Brief Survey of Automatic Methods for Author Name Disambiguation (Ferreira et al)

- Automatic Methods
  - **Author Grouping** - Groups authors using similarity based on references
    - Pre-Defined functions - such as TFIDF
    - Learned similarity function - similar to above with training data
    - Graph based similarity function – Similar to social network analysis
    - Uses clustering techniques
  - **Author Assignment** – Create model of author then assign items to them
    - Classification – Supervised machine learning requires large training set
    - Clustering
  - **Group by alternative info**
    - Web-Searches etc. on authors

Papers

A Brief Survey of Automatic Methods for Author Name Disambiguation (Ferreira et al)

- **Very little data in the citations**
  - Needs additional data
- Ambiguous cases
- Citations with errors
- Different Knowledge areas
- **Author profile changes**

Papers

Name Disambiguation in Author Citations using a K-way Spectral Clustering Method (Han et al)

Uses predefined functions

- TFIDF = tf(t,d) * idf (t,D)
  - Term Frequency Inverse Document Frequency
  - The more documents the term appears in the less important it is
- NTF ntf(i, d) = freq(i, d)/max(freq(i, d)) freq(i, d)
  - Normalized Term Frequency
  - Uses author features (co-authors, titles, venues)
- Uses K-Way Spectral clustering
  - More features improves accuracy of results

Papers

Fast Author Name Disambiguation in CiteSeer

- CiteSeer
  - Repository of papers
- Uses DBSCAN for clustering
- Proposes scaling issues with Han due to choosing K
- Uses support vector machine for distance calculation
- Seems to fall into the Author Assignment using NTF category
- Lots of citations

Papers

Efficient Topic-based Unsupervised Name Disambiguation (Song et al)

- Proposes a two stage approach
  - Create a topic based model
    - Richard Ernst (Infosec) – Richard Ernst (Chemistry)
  - Topics are then used as features
- Purports to outperform DBSCAN methods (Huang)
- Defines categories of classification as Supervised and Unsupervised
  - Large Scale requires Unsupervised
- The relationships between documents, names and words are important
  - Policy Enforcement (Infosec) - Nuclear magnetic resonance spectroscopy (Chemistry)

Applicaton

Data Source - http://www.securityfocus.com/bid/64078

info | discussion | exploit | solution | references

Google Chrome Prior to 31.0.1650.63 Multiple Security Vulnerabilities

Bugtraq ID:	64078
Class:	Unknown
CVE:	CVE-2013-6637 CVE-2013-6638 CVE-2013-6639 CVE-2013-6634 CVE-2013-6636 CVE-2013-6640
Remote:	Yes
Local:	No
Published:	Dec 04 2013 12:00AM
Updated:	Mar 07 2014 01:03AM
Credit:	Andrey Labunets,cloudfuzzer,Bas Venis, and Jakob Kummerow
Vulnerable:	Google Chrome 16.0.912 75 Google Chrome 15.0.874 102 Google Chrome 3.0.195 .21 Google Chrome 0.3.154 9 Google Chrome 9.0.597.94

Credit for discovery of the vulnerability

Goal

- Assign categorical enumeration to each CVE based upon value of **Credit**
- Categories
  - Anonymous
  - HackerAlias
  - NamedResearcher
  - ResearchLab
  - University
  - Unknown
  - Vendor

The programmer approach

- Supervised learning
  - Create regex of text corresponding to categories
    ResearchLab/GoodCriteria.txt
    - Gmb[hH]
    - BugSec LTD
    - Pentesting
    - RedTeam
    - Vupen
    - Day Initiave
    - iDefense Labs
    - Google Security Team
- Create regex for negative cases
  - If a category matches Research Lab – Lab criteria becomes filter for University.

Current Data Processing
Assigning Categoricals

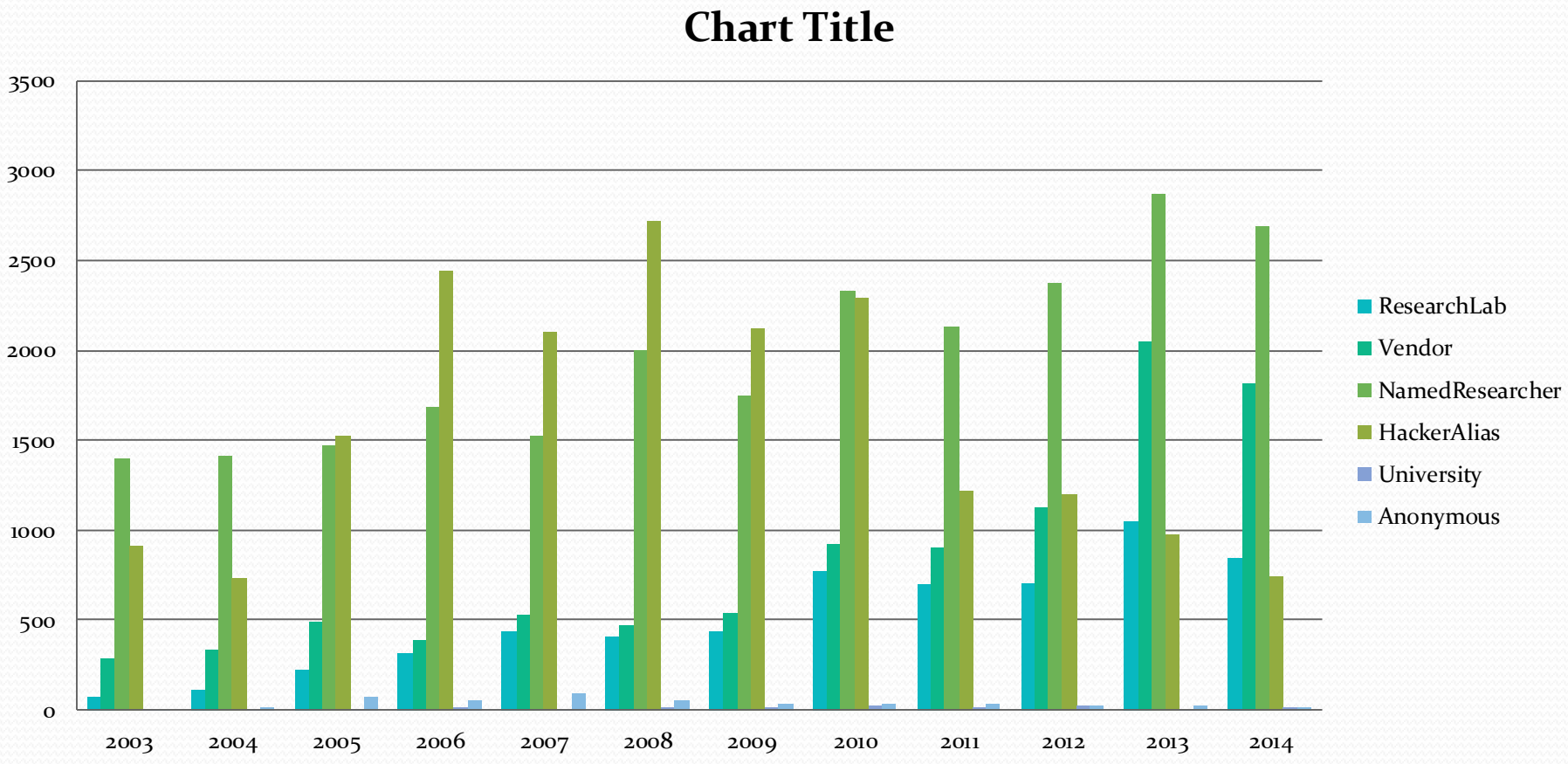
```
#!/bin/sh

LAST_CATEGORY=""
discovery=$1
BASEDIR=`dirname $0`
for categorical in ResearchLab Vendor University Anonymous NamedResearcher Unknown HackerAlias;do
  echo "Processing $categorical"
  rm $BASEDIR/all.txt
  if [ $LAST_CATEGORY != "" ];then
    rm $BASEDIR/${categorical}/BadCriteria.txt
  fi
  cp $BASEDIR/$LAST_CATEGORY/BadCriteria.txt $BASEDIR/${categorical}/
  cat $BASEDIR/$LAST_CATEGORY/GoodCriteria.txt >> $BASEDIR/${categorical}/BadCriteria.txt
  rm $BASEDIR/${categorical}/${categorical}_filtered.txt
  categorize $discovery $categorical
  LAST_CATEGORY=$categorical
done
rm $BASEDIR/all_updates.sql
rm $BASEDIR/all_filtered-categorized.txt
rm $BASEDIR/all_filtered-uncategorized.txt
for categorical in ResearchLab Vendor University Anonymous NamedResearcher Unknown HackerAlias;do
  cat $BASEDIR/${categorical}/${categorical}_filtered.txt |awk -v CA=${categorical} '{print  "UPDATE vulnerabilityInfo_tbl
SET author_category=\x27"CA"\x27 WHERE vulnerabilityId=\x27" $1 "\x27;"}' >>  $BASEDIR/all_updates.sql
  cat $BASEDIR/${categorical}/${categorical}_filtered.txt |sed "s/$/\t${categorical}/" >>  $BASEDIR/all_filtered-
categorized.txt
  cat $BASEDIR/${categorical}/${categorical}_filtered.txt  >>  $BASEDIR/all_filtered-uncategorized.txt
done
```

Current Processed Data

SecurityFocus ID	CVE ID	Disclosure Date	Author	Author Category
64078	CVE-2013-6637	12/4/2013	Andrey Labunets,cloudfuzzer,Bas Venis, and Jakob Kummerow	ResearchLab
64078	CVE-2013-6638	12/4/2013	Andrey Labunets,cloudfuzzer,Bas Venis, and Jakob Kummerow	ResearchLab
64078	CVE-2013-6639	12/4/2013	Andrey Labunets,cloudfuzzer,Bas Venis, and Jakob Kummerow	ResearchLab
64078	CVE-2013-6634	12/4/2013	Andrey Labunets,cloudfuzzer,Bas Venis, and Jakob Kummerow	ResearchLab
64078	CVE-2013-6636	12/4/2013	Andrey Labunets,cloudfuzzer,Bas Venis, and Jakob Kummerow	ResearchLab
64078	CVE-2013-6640	12/4/2013	Andrey Labunets,cloudfuzzer,Bas Venis, and Jakob Kummerow	ResearchLab
71401	CVE-2014-3809	12/1/2014	Stephan Rickauer	NamedResearcher
71402	CVE-2014-8104	12/1/2014	Dragana Damjanovic	NamedResearcher
71403		12/2/2014	LiquidWorm	HackerAlias
71404	CVE-2014-5446	12/1/2014	Pedro Ribeiro	NamedResearcher
71404	CVE-2014-5445	12/1/2014	Pedro Ribeiro	NamedResearcher
71405		12/2/2014	LiquidWorm	HackerAlias

Data Analytics



Current Data Processing

- Primitive
- Fast to develop
- Scales poorly
- Easy to understand
- Requires new approach Let's discuss
- Luckily rather small data set (68K)

New Approach

- Extend Song approach
- Will require parsing **Credit** field
- Likely require new relationship in DB
  - Name is assigned category
    - Andrey Lamberts = Named Researcher
    - Cloudfuzzer = Hacker Alias
- Topics should align with affected products
- Should get info on related researchers
  - Who does Cloudfuzzer regularly work with?
- Should get new info on how individuals work
  - Does Cloudfuzzer specialize in Chrome or Buffer overflows?
- Should get new info on

Issues

- Still need to determine if elaich is individual's name or a stylized Hacker Alias
- Still need disambiguate Microsoft (vendor) from Cloudfuzzer (Hacker Alias)
- How do you categorize marc@EEYE.COM
  - Individual (Mark ?)
  - Eeye ResearchLab?
  - Punt / Guess?