Welcome to the class of Advanced Topics in Information Retrieval !

Min ZHANG （张敏）

z-m@tsinghua.edu.cn

Tea Time
Dangerous Ransomware

Min Zhang

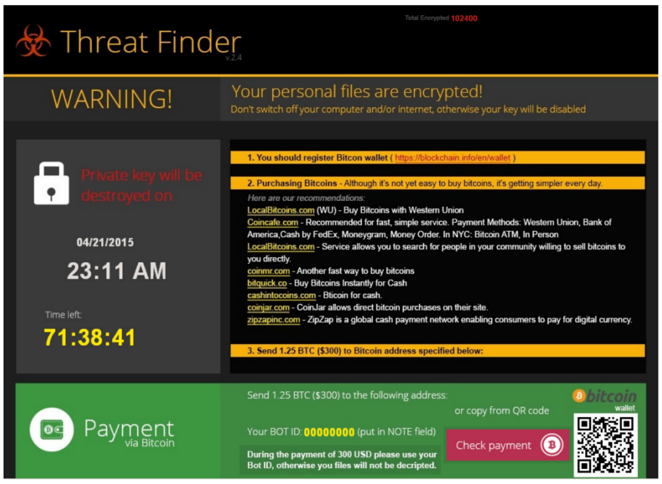
z-m@tsinghua.edu.cn

Dangerous Ransomware

Types

- Locker Ransomware (computer locker)
  - Denies access to the computer or device
  - Leaves the underlying system and files untouched
  - Less effective at extracting ransom compared with crypto ransomware
- Crypto Ransomware (data locker)
  - Prevents access to files or data
  - Finds and encrypts valuable data using 2048 or 4096-bit RSA keys
  - Encrypted files are unusable unless decryption key is obtained

Crypto Ransomware Demand Screen



Total Encrypted 102400

Threat Finder v.2.4

WARNING!

Your personal files are encrypted!

Don't switch off your computer and/or internet, otherwise your key will be disabled

Private key will be destroyed on

04/21/2015

23:11 AM

Time left:

71:38:41

1. You should register Bitcon wallet (https://blockchain.info/en/wallet)

2. Purchasing Bitcoins - Although it's not yet easy to buy bitcoins, it's getting simpler every day.

Here are our recommendations:

LocalBitcoins.com (WU) - Buy Bitcoins with Western Union
Coincafe.com - Recommended for fast, simple service. Payment Methods: Western Union, Bank of America,Cash by FedEx, Moneygram, Money Order. In NYC: Bitcoin ATM, In Person
LocalBitcoins.com - Service allows you to search for people in your community willing to sell bitcoins to you directly.
coinmr.com - Another fast way to buy bitcoins
bitquick.co - Buy Bitcoins Instantly for Cash
cashintocoins.com - Bticoin for cash.
coinjar.com - CoinJar allows direct bitcoin purchases on their site.
zipzapinc.com - ZipZap is a global cash payment network enabling consumers to pay for digital currency.

3. Send 1.25 BTC ($300) to Bitcoin address specified below:

Payment via Bitcoin

Send 1.25 BTC ($300) to the following address:

or copy from QR code

Your BOT ID: 00000000 (put in NOTE field)

During the payment of 300 USD please use your Bot ID, otherwise you files will not be decripted.

Key Figures

- Emerged in 2013 with CryptoLocker
  - Android variant was reported in 2014
  - 17% of the infections in 2015 were on Android devices
  - In March 2016, an Apple Mac variant was found
- Ransomware programs were detected on 753,684 computers in 2015
  - 179,209 computers were targeted by encryption ransomware.
- “The ransomware is that good… To be honest, we often advise people just to pay the ransom.”
  - Joseph Bonavolonta, an assistant special agent with the FBI, at Boston's Cyber Security Summit 2015

Paralyzed by Ransomware

- On February 2016 a hospital in Los Angeles was hit by ransomware, leaving doctors unable to access critical patient data for more than a week
- It was reported that all patient-record history and hospital email archives were encrypted by the malware
- Cybercriminals asked for $3.6 million in Bitcoin for the decryption key
- However, it is uncertain whether the hospital has paid the ransom or not

Paralyzed by Ransomware

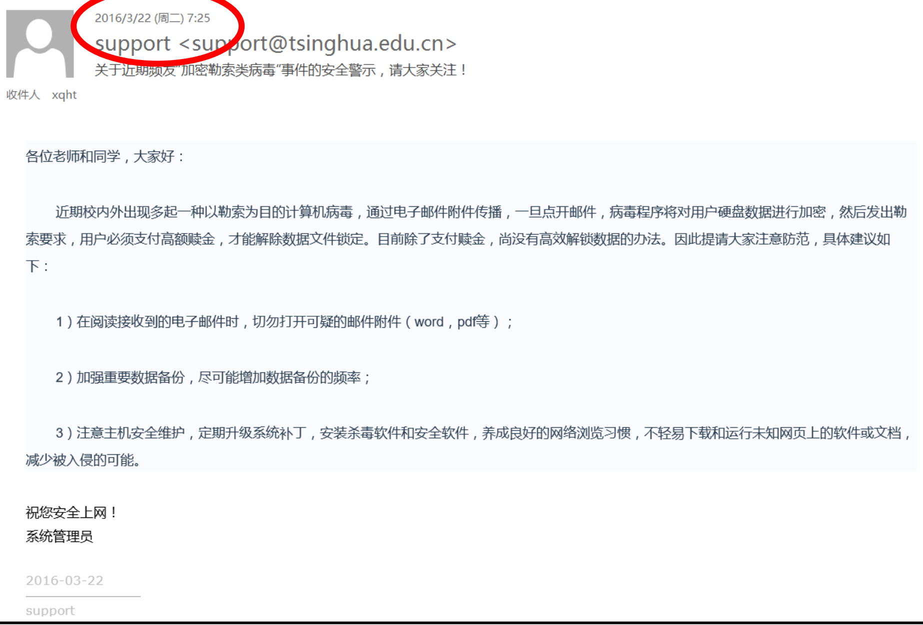
Police Pay Ransomware Demand

- Computer system of Tewksbury Police Department, Massachusetts was infected early March with ransomware
- To keep their computer files from being destroyed Chief Timothy Sheehana authorised the $550 ransom demand in bitcoin
- It is believed a communal network user accidentally downloaded the malware, which then encrypted all the computer data, holding it for ransom.

Prevention

- Backup your files regularly
- Apply software patches as soon as they become available
  - Some ransomware arrive via vulnerability exploits.
- Bookmark trusted websites and access these websites via bookmarks
- Download email attachments only from trusted sources
- Scan your system regularly with anti-malware

Prevention



2016/3/22 (周二) 7:25

support <support@tsinghua.edu.cn>

关于近期频发"加密勒索类病毒"事件的安全警示，请大家关注！

收件人　xqht

各位老师和同学，大家好：

近期校内外出现多起一种以勒索为目的计算机病毒，通过电子邮件附件传播，一旦点开邮件，病毒程序将对用户硬盘数据进行加密，然后发出勒索要求，用户必须支付高额赎金，才能解除数据文件锁定。目前除了支付赎金，尚没有高效解锁数据的办法。因此提请大家注意防范，具体建议如下：

1）在阅读接收到的电子邮件时，切勿打开可疑的邮件附件（word，pdf等）；

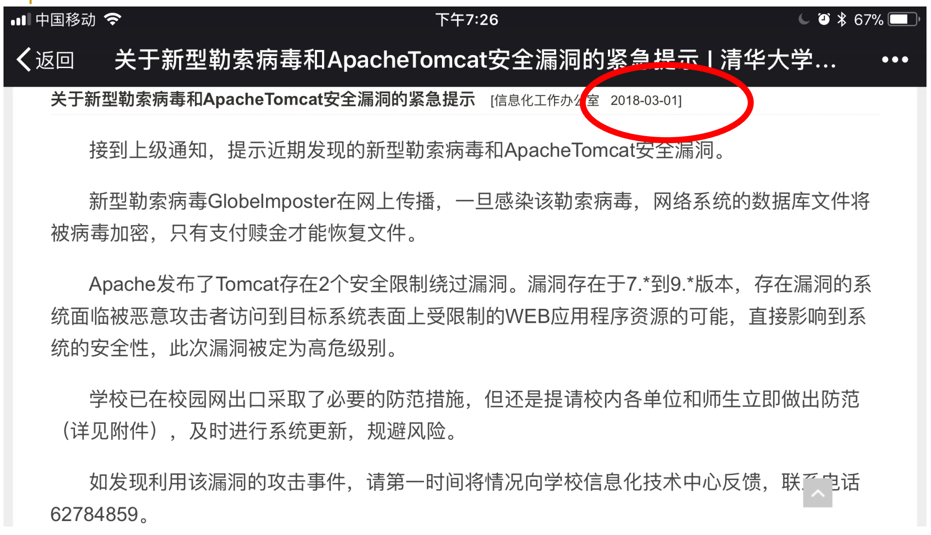
2）加强重要数据备份，尽可能增加数据备份的频率；

3）注意主机安全维护，定期升级系统补丁，安装杀毒软件和安全软件，养成良好的网络浏览习惯，不轻易下载和运行未知网页上的软件或文档，减少被入侵的可能。

祝您安全上网！
系统管理员

2016-03-22

support



中国移动　下午7:26　67%

返回　关于新型勒索病毒和ApacheTomcat安全漏洞的紧急提示 | 清华大学...

关于新型勒索病毒和ApacheTomcat安全漏洞的紧急提示　[信息化工作办公室　2018-03-01]

接到上级通知，提示近期发现的新型勒索病毒和ApacheTomcat安全漏洞。

新型勒索病毒GlobeImposter在网上传播，一旦感染该勒索病毒，网络系统的数据库文件将被病毒加密，只有支付赎金才能恢复文件。

Apache发布了Tomcat存在2个安全限制绕过漏洞。漏洞存在于7.*到9.*版本，存在漏洞的系统面临被恶意攻击者访问到目标系统表面上受限制的WEB应用程序资源的可能，直接影响到系统的安全性，此次漏洞被定为高危级别。

学校已在校园网出口采取了必要的防范措施，但还是提请校内各单位和师生立即做出防范（详见附件），及时进行系统更新，规避风险。

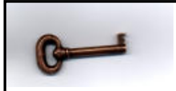
如发现利用该漏洞的攻击事件，请第一时间将情况向学校信息化技术中心反馈，联系电话62784859。

March 10, 2019

病毒名称：GlobeImposter3.0 变种
病毒性质：勒索病毒
影响范围：多省份出现医院大规模爆发，有全国爆发趋势
危害等级：高危

紧急公告

近日，深信服安全团队发现GlobeImposter勒索病毒3.0变种再次席卷全国各地医院，受影响的系统，数据库文件被加密破坏，病毒将加密后的文件后缀改以*4444结尾，并要求用户通过邮件沟通赎金跟解密密钥等。目前GlobeImposter 3.0已在多个省份形成规模爆发趋势，深信服再次发布紧急预警，建议全国各医院做好安全防护，警惕GlobeImposter 勒索，各医院可联系当地深信服技术人员或市场人员，快速获取病毒应急响应支撑。

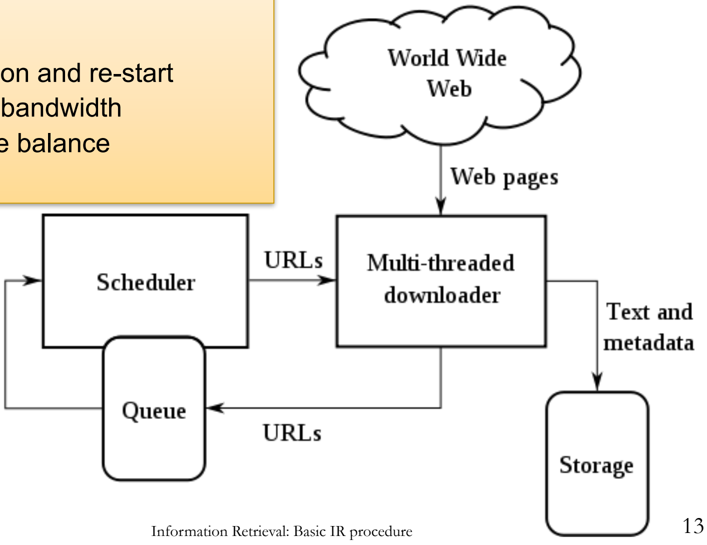


Brief introduction to IR fundamentals – II

Crawler (cont.) and Indexing

Review: Web Crawler Architecture

- Breadth-first or Depth-first?
- Priority
- Timeout
- Interruption and re-start
- Network bandwidth
- Resource balance
- ……



Review:

"It is fairly easy to build a slow crawler that downloads a few pages per second for a short period of time, building a high-performance system that can download hundreds of millions of pages over several weeks presents a number of challenges in system design, I/O and network efficiency, and robustness and manageability."

Eichmann, D. (1994). The RBSE spider: balancing effective search against Web load. In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.

What any crawler *must* do

- Be Polite: Respect implicit and explicit politeness considerations for a website
  - Explicit politeness: specifications from webmasters on what portions of site can be crawled
    - robots.txt ➡

Robots.txt

- Protocol for giving spiders (“robots”) limited access to a website, originally from 1994
  - www.robotstxt.org/orig.html
- Website announces its request on what can(not) be crawled
  - For a URL, create a file **`URL/robots.txt`**
  - This file specifies access restrictions
- Example: # robots.txt

```
User-agent: *
Disallow: /cyberworld/map/ # This is an infinite virtual URL space

# Cybermapper knows where to go.
User-agent: cybermapper
Disallow:
```

What any crawler *must* do

- Be Polite: Respect implicit and explicit politeness considerations for a website
  - Explicit politeness: specifications from webmasters on what portions of site can be crawled
    - robots.txt
  - Implicit politeness: even with no specification, avoid hitting any site too often
- Be Robust: Be immune to spider traps and other malicious behavior from web servers

What any crawler *should* do

- Be capable of distributed operation: designed to run on multiple distributed machines (data coverage)
- Be scalable: designed to increase the crawl rate by adding more machines
- Performance/efficiency: permit full use of available processing and network resources (minimize server loads)
- Fetch pages of "higher quality" first (index "good" pages, no duplicates)
- Continuous operation: Continue fetching fresh copies of a previously fetched page (always keep "fresh" pages)
- Extensible: Adapt to new data formats, protocols

Current Challenging Topics

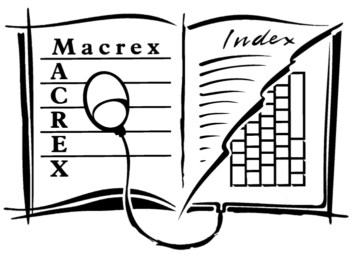
- Focused crawler
  - Crawler for vertical domain, Topic-specific crawler, …
- Dynamic pages, new UI techniques
- Crawling the deep / hidden / invisible web
- Crawling the real time data (e.g. Twitter, Weibo, …)
- ……

Some recommended open source crawlers

- Heritrix
  - https://webarchive.jira.com/wiki/spaces/Heritrix/
  - The Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project.
- Nutch (A web search engine software, not only a crawler)
  - Crawler
  - Indexer and search system
  - http://nutch.apache.org/
- GNU Wget
  - A free software package for retrieving files using HTTP, HTTPS and FTP
  - http://www.gnu.org/software/wget/
- Crawler4j
  - Open source web crawler for java, multithread
  - https://github.com/yasserg/crawler4j

Outline

- Basic IR procedure
  - Data acquisition
  - Indexing
  - Ranking
  - System evaluation



Why indexing?

- Search and fast access to the content (in Nature Language, multimedia, etc.)
- Possible approaches:
  1. String matching (linear search in documents)
     - Slow
     - Difficult to improve
  2. Indexing
     - Fast
     - Flexible to further improvement

Why indexing?

- Goal = Find the important concepts and create an internal representation
- Factors to consider:
  - Accuracy to represent concepts (semantics)
  - Facility for computer to manipulate
- What is the best representation of concepts?

  Keywords are a simple and effective way,

  but **not the most precise**.

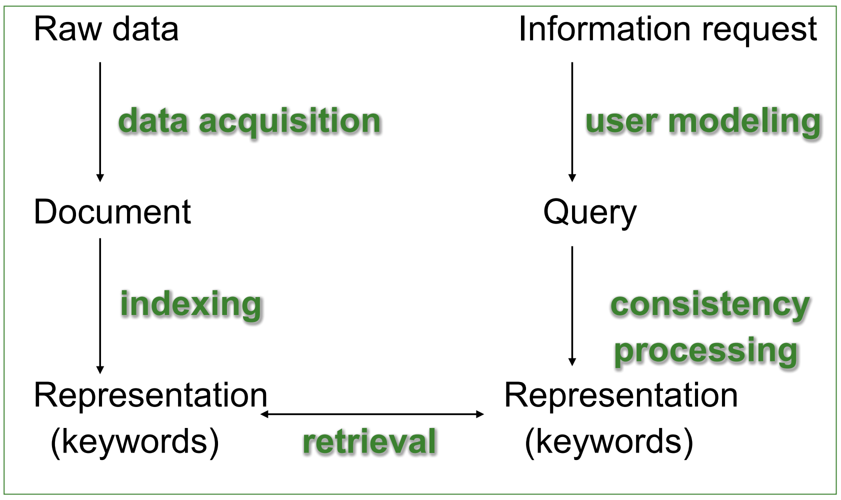
What to index?

- What is the best representation of concepts?

  Keywords are a simple and effective way,

  but not the most precise.
- Not all information need can be expressed by one or several keywords
  - e.g. “to be or not to be”   in "Hamlet"
- Not all terms are useful.
  - e.g. “What is the URL of Google search engine”
  - e.g. “I met a computer problem. Error code : 0x00041”
- But it is still the most dominant representation in the state-of-art IR systems

Indexing-based IR



Result of indexing – inverted index

- Each document corresponds to a set of weighted keywords:

  $D_1 \rightarrow \{(t_1, w_1), (t_2, w_2), \ldots\}$

- Inverted file:

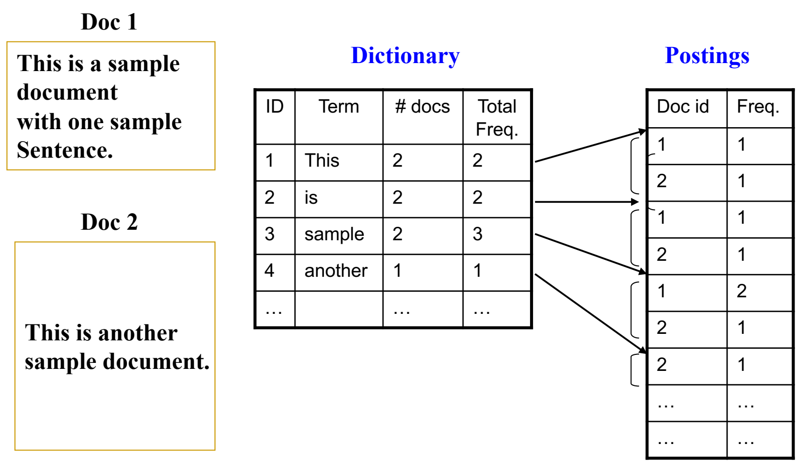
  $t_1 \rightarrow \{(D_1, w_{11}, [p_{11}]), (D_1, w_{12}, [p_{12}]), \ldots, (D_i, w_{i1}, [p_{i1}]), \ldots\ldots\}$

  Inverted file is used during retrieval for higher efficiency.

p: the term position in the document

D: document, t: selected term, w: term weights

Inverted Index Example



Doc 1

This is a sample document with one sample Sentence.

Doc 2

This is another sample document.

Dictionary

ID	Term	# docs	Total Freq.
1	This	2	2
2	is	2	2
3	sample	2	3
4	another	1	1
…		…	…

Postings

Doc id	Freq.
1	1
2	1
1	1
2	1
1	2
2	1
2	1
…	…
…	…

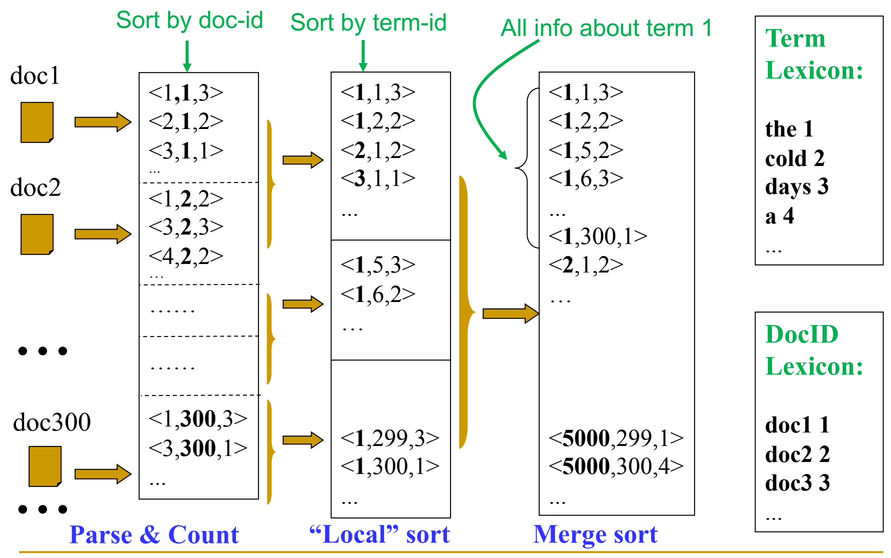
Data Structures for Inverted Index

- Dictionary: modest size
  - Needs fast random access
  - Preferred to be in memory
  - Hash table, B-tree, Trie, …
- Postings: huge
  - Sequential access is expected
  - Can stay on disk
  - May contain docID, term freq., term pos, etc
  - Compression is desirable
    - Exploit skewed frequency distribution: fewer bits for small (high frequency) integers

Constructing Inverted Index

- The major difficulty is to build a huge index with limited memory
- Memory-based methods: not usable for large collections
- Sort-based methods:
  - Step 1: collect local (termID, docID, freq) tuples
  - Step 2: sort local tuples (to make "runs")
  - Step 3: pair-wise merge runs
  - Step 4: output inverted file

Sort-based Inversion



Term selection (before indexing)

- In many cases, you don't need to index every term.
- Filter out stopwords/function words (that are meaningless) e.g. *of, in, about, I*, …
- Remove insignificant differences by word stemming (or by applying some morphological transformation - lemmatization), in order to increase recall

  transforming, transforms, transformed → transform

  transformation → transform   computer → comput

  (crucial to choose stemming rules to avoid much noise)

DISCUSSIONS

Discussion 1: dealing with phrases

- *How can we represent* ***phrases in the index****?*
  - e.g. “to be or not to be”
  - “清华大学”
- *How can we deal with* ***position*** *information if we use* ***multi-granularity index****?*
  - E.g. “To be or not to be: that is a question”.

Discussion 2:

- *How many ways can you find to* ***prevent*** *your website from* ***being indexed*** *by search engines?*