



TABLUS

The Security Role for Content Analysis

Jim Nisbet
Founder, Tablus, Inc.

November 17, 2004

It's your business to know.™

About Us



- Tablus is a 3 year old company that delivers solutions to provide visibility to sensitive information leaving the network
- Our background
 - low-level packet capture and protocol analysis
 - content analysis, content classification
 - database management

It's your business to know.™

What is Content Analysis?



- Techniques for extracting and classifying data
 - Simplistic (deterministic)
 - Matching specific keywords or patterns
 - Matching meta data in specific file formats
 - Algorithmic
 - Statistical analysis
 - Language based analysis
 - Document clustering

It's your business to know.™

What Does Content Analysis have to do with Security?



- Companies are increasingly focused on protecting their sensitive information
 - Customer lists, SSNs (PII, PHI)
 - Trade secrets, internal documents
 - Proprietary information -- source code, etc.

The challenge is to be able to accurately classify data as sensitive

It's your business to know.™

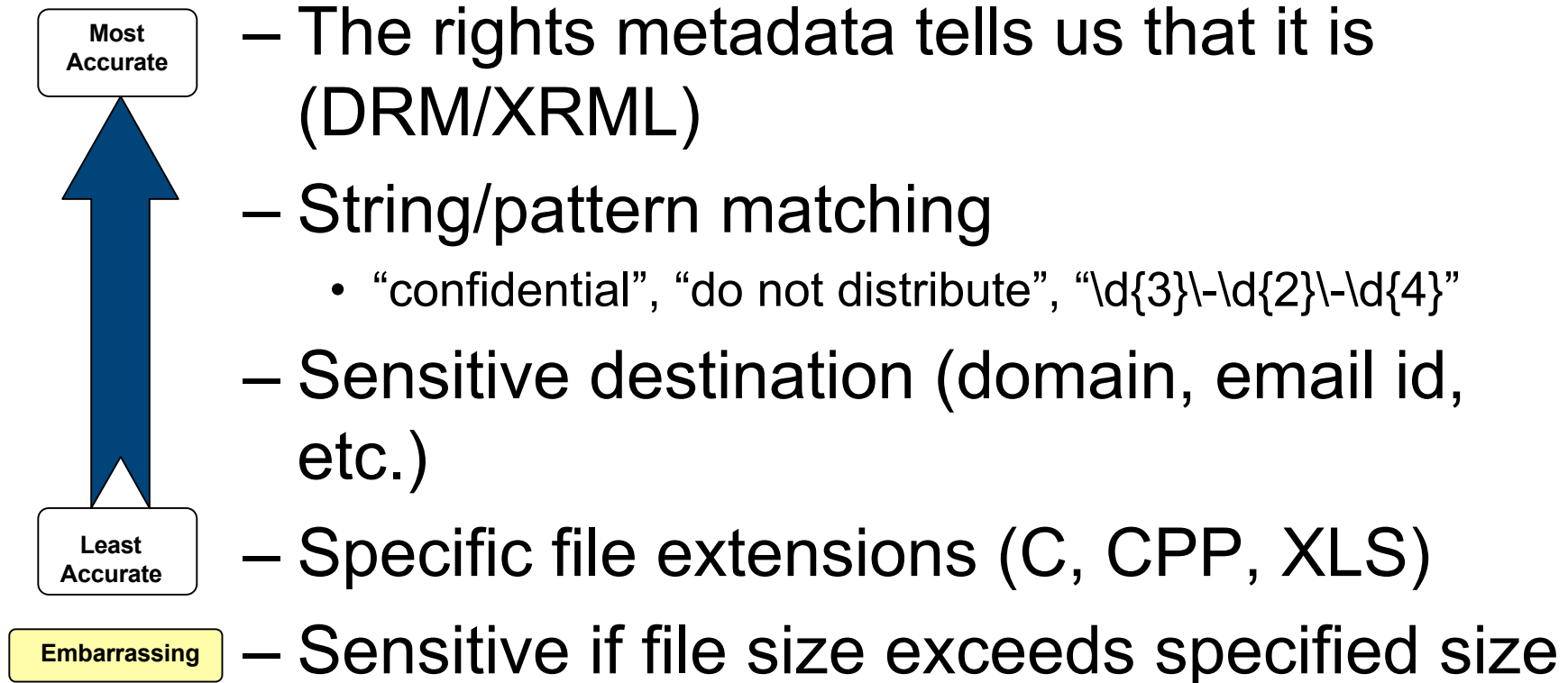
Detecting Sensitive Information



- Verify that copies of sensitive information are adequately protected
- Remove unneeded copies of sensitive information
- Monitor the network for the transmission of sensitive information

It's your business to know.™

Real World Ways Mail Filters Identify Information as Sensitive...



It's your business to know.™

Types of Machine Classification Techniques



- Linguistic
 - Based on the text or language
 - Semantic analysis
 - Based on the meaning of the words
- Bayesian Inference
 - Based on probability theory
- Statistical/AI
 - Neural net
- Hybrids



Lexical Resources
(e.g. wordnet)



Training Set

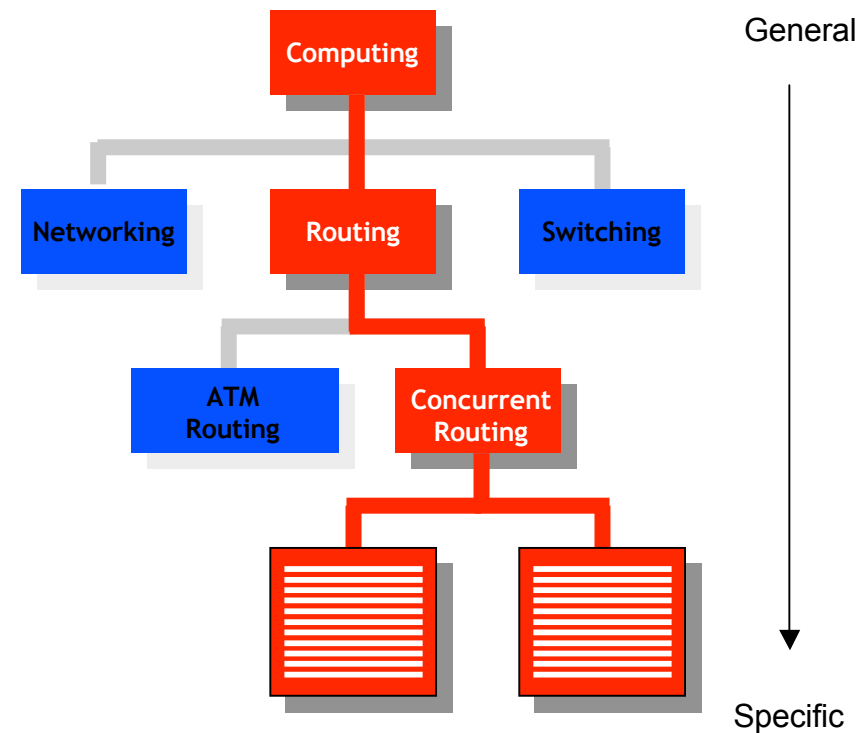
It's your business to know.™

Taxonomic Classification



- Taxonomic classification can tell you what subjects are likely discussed in a document

- Would allow us to say “documents that talk about concurrent routing are sensitive”

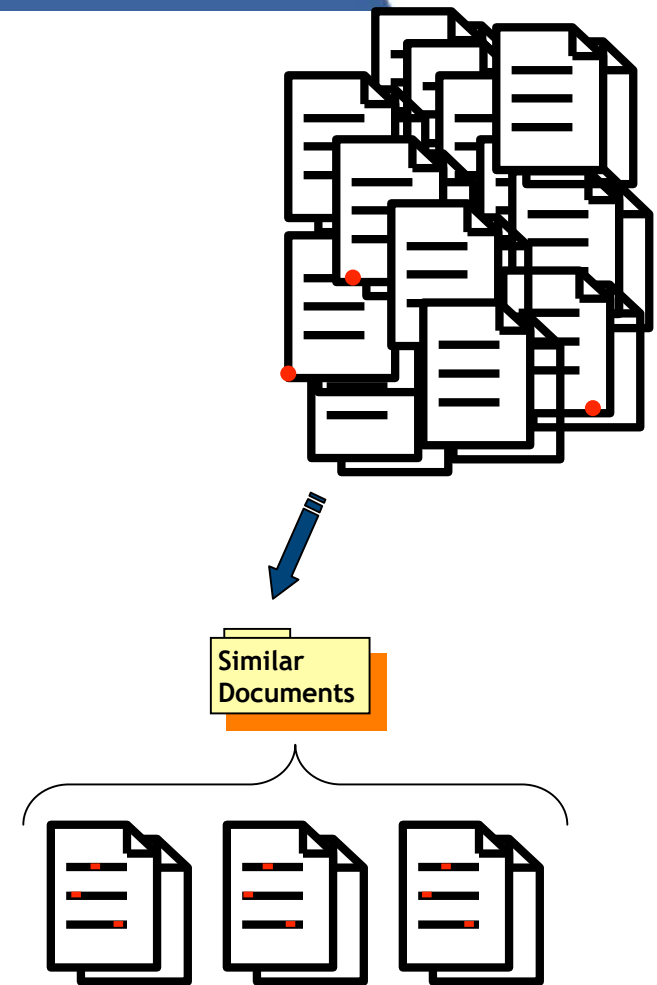


It's your business to know.™

Document Clustering



- Documents with similar characteristics grouped together
 - Based on language
 - Based on statistical information
 - Based on meta data

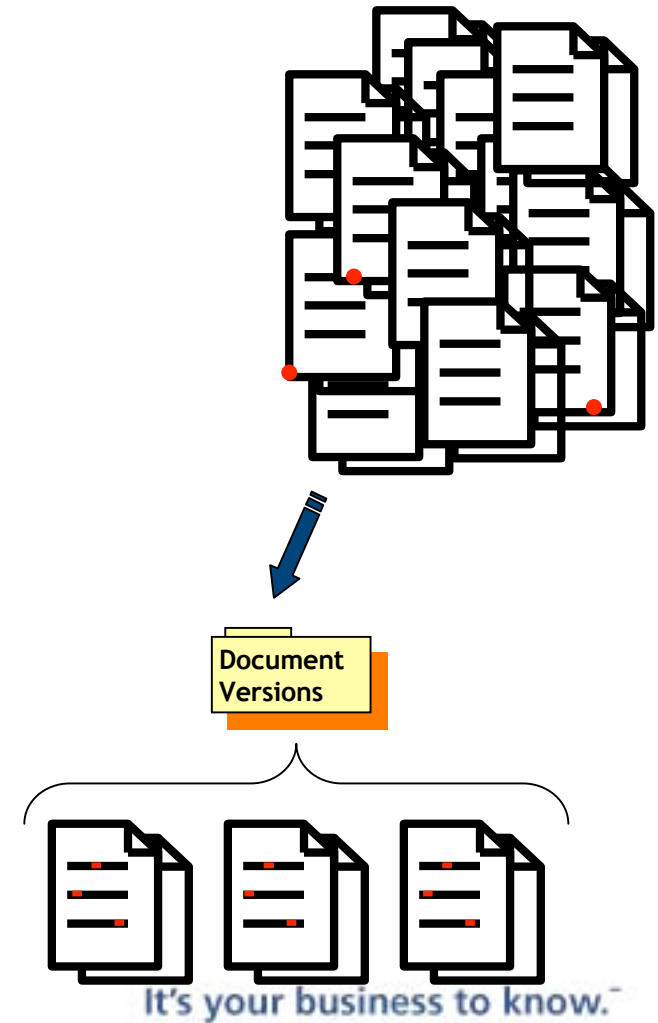


It's your business to know.™

Derivative Work Analysis



- Derivative work analysis means looking for overwhelming similarity in lexicon or form



It's your business to know.™

Vector Space Analysis



- Document represented by a vector of terms
 - Words (or word stems)
 - Phrases (e.g. computer science)
 - Removes words on “stop list”
 - Documents aren’t about “the”
- Often assumed that terms are uncorrelated
- Correlations between term vectors implies a similarity between documents

It's your business to know.™

Document Vectors



- Documents are represented as “bags of words”
- Represented as vectors when used computationally
 - Has direction and magnitude
 - Each vector holds a place for **every** term in the collection
 - Therefore, most vectors are sparse

It's your business to know.™

The Training Set: Processing Existing Sensitive Content



- We can gain accuracy if we can provide
 - A representative set of sensitive content
- Negative training set (“outliers”)
 - Similar content that is not sensitive
- Learning systems
 - Supervised vs. Unsupervised learning

It's your business to know.™

Spam vs. Sensitive Content



Spam detection and sensitive content detection have significant differences.

It's your business to know.™

Detecting Sensitive Information vs. Spam Detection



- **Similarities**

- Language analysis can be used to classify
- Meta data may exist to help with detection

- **Differences**

- Your Spam and my Spam are the same
- Your sensitive information and mine are different
- Sensitive information is often time sensitive; Spam is not
- Spammers try to make Spam not look like Spam; sensitive information is not obfuscated

It's your business to know.™

**HOW DO YOU KNOW WHAT
IS SENSITIVE?**



ASSET CLASSIFICATION

It's your business to know.™

ISO 17799: Code of Practice for Information Security Management



- Security Policy
- System Access Control
- Computer & Operations Management
- System Development and Maintenance
- Physical and Environmental Security
- Compliance
- Personnel Security
- Asset Classification and Control
- Business Continuity Management

It's your business to know.™

ISO 17799: Code of Practice for Information Security Management



- Security Policy
- System Access Control
- Computer & Operations Management
- System Development and Maintenance
- Physical and Environmental Security
- Compliance
- Personnel Security
- **Asset Classification and Control**
- Business Continuity Management

It's your business to know.™

ISO 17799: Asset Classification



A suggested schema for classification:

- **Confidentiality** — can the information be freely distributed?
- **Value** — is this costly to replace?
- **Time** — will its confidentiality status change over time
- **Access rights** — who will have access to this information?
- **Retention** — when can this information be destroyed?

It's your business to know.™

ISO 17799: Asset Classification



A suggested schema for classification:

- **Confidentiality** — can the information be freely distributed?
- **Value** — is this costly to replace?
- **Time** — will its confidentiality status change over time
- **Access rights** — who will have access to this information?
- **Retention** — when can this information be destroyed?

It's your business to know.™

ISO 17799: Confidentiality



Suggested levels of confidentiality*:

- **Authentication data** — internal security information
- **Restricted** — never leaves organization
- **Confidential Plus** — financial information
- **Confidential** — restricted to specific roles within the company
- **Internal** — restricted to company employees
- **Public** — publicly accessible information

* Actual classification schema used by a very large financial institution

It's your business to know.™

INFORMATION RETRIEVAL (IR)



IT'S ALL ABOUT THE ACCURACY

It's your business to know.™

Quick Overview of Search and IR Technologies



- In IR solutions, when we talk about “accuracy” we often talk about:
 - Precision
 - Proportion of retrieved material actually relevant
 - Recall
 - Proportion of relevant material actually retrieved

It's your business to know.™

Is Internet Search really about “search”?



- Finding all occurrences is not too helpful
- Ranking the relevance of the results is key

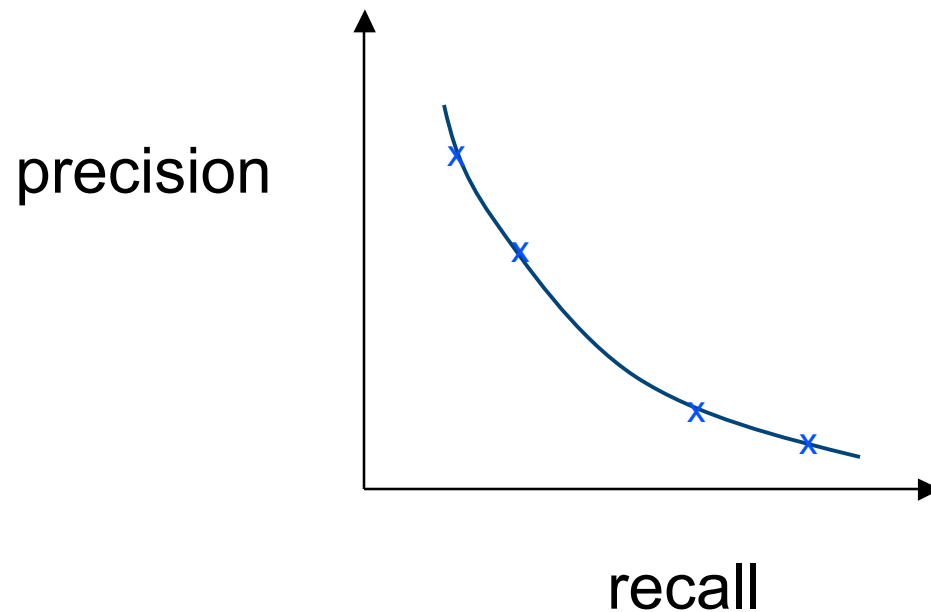
-
- If you lower the Recall you will generally increase the Precision but you run the risk of missing something you want.

It's your business to know.™

Precision/Recall Curves



- There is a tradeoff between Precision and Recall
- So measure Precision at different levels of Recall



It's your business to know.™

TREC – NIST Text Retrieval Competition



- Text REtrieval Conference/Competition
 - Run by NIST (National Institute of Standards & Technology)
- Collection: 3 Gigabytes, >1 Million Docs
 - Newswire & full text news (AP, WSJ, Ziff)
 - Government documents (federal register)
- Queries + Relevance Judgments
 - Queries devised and judged by Information Specialists
- Competition
 - Various research and commercial groups compete
 - Results judged on precision and recall

It's your business to know.™

An Example of Poor Precision...



The screenshot shows a Google search interface with the search term "plumber menlo park" entered in the search box. The search results are displayed under the "Web" tab, showing 10 results out of approximately 2,300. The results include:

- June 2000 : Informix & amp; XML**
... Chief **Plumber**. ... INFORMIX Software Inc. , 4100 Bohannon Drive **Menlo Park** California 94025 USA Phone: 650 926 6300 Fax: 510 628 3951 email: paul.brown@informix.com ...
www.infoloom.com/gcaconfs/WEB/paris2000/S35-04.HTM - 30k - [Cached](#) - [Similar pages](#)
- Deaths**
... He was a retired **plumber** and laborer and a member of the Masonic Lodge. ... of Modesto and Betty Glass of Palo Alto; a brother, James Linel of **Menlo Park**; a sister ...
www.paloaltoonline.com/weekly/morgue/community_pulse/1998_Oct_28.OBITS28.html - 7k - [Cached](#) - [Similar pages](#)
- News Digest**
... A **plumber** working at a construction site owned by the city of **Menlo Park** discovered an active pipe bomb there on Aug. 29, **Menlo Park** police said. ...
www.paloaltoonline.com/weekly/morgue/news/2000_Sep_6.DIGEST6.html - 4k - [Cached](#) - [Similar pages](#)
[[More results from www.paloaltoonline.com](#)]
- Menlo Park election: Speer says he'll return contribution**
... Nicholas Jellins: **Plumber** & Steamfitters ... Al and Sarah Filice, (he) president and CEO of Filice Lansford Development, and president of the **Menlo Park** Chamber of ...
www.almanacnews.com/morgue/2002/2002_10_16.mpfinance.html - 8k - [Cached](#) - [Similar pages](#)
- Menlo election: Candidates are neck and neck on fund-raising**
... 1. Each received \$750 from the **Plumber** and Steamfitters Union Local 467; and \$300

The browser's address bar shows the URL: http://www.google.com/search?hl=en&lr=&q=related:www.almanacnews.com/morgue/2002/2002_10_16

Precision and Ambiguity



The screenshot shows a Google search interface with the query "unicycle repair instructions". The search results are displayed under the "Web" tab. The first result is "Frequently Asked Questions on Unicycling" from unicycling.org. The second result is "Unicycle for Smalltalk" from unity-software.com. The third result is "Unicycle for Smalltalk" from unity-software.com. The fourth result is "Goudurix ::" from goudurix.com. The fifth result is "Unicyclist Community Forums - KH 20 / KH 24 hub moving?" from unicyclist.com. A yellow callout box highlights the first two results, stating: "For search, people often automatically disambiguate the results without ever realizing that the search engine could not do this for them. If the answer you are looking for is in the first five results returned it may not matter if we can not correctly rank the closest match as #1." A yellow box highlights the fifth result, which contains the text: "Unicyclist Community Forums - KH 20 / KH 24 hub moving? ... You've got three choices 1) Contact Unicycle.com and find out if they will replace or repair the wheel 2) Send the wheel to Darren Bedford to fix 3) Fix it ...".

Google Web Images Groups News Froogle Desktop more »

unicycle repair instructions Search Advanced Search Preferences

Web Results 1 - 10 of about 923 for **unicycle repair instructions**. (0.17 seconds)

[Frequently Asked Questions on Unicycling](#)
... You can buy kits for these from the **Unicycle** Factory. Im the BC wheel. ... 4. Maintenance and **Repair**. 4.1 How do u
www.unicycling.org/unicycling/faq.html - 78k - 19 Oct 2004 -

[Unicycle for Smalltalk](#)
... file. **UniCycle** V2.3. **UniCycle** V2.1 US21FIX.EXE (120K
Enclose user values in double quotes for CSV export. Fix ...
www.unity-software.com/html/support/ufs/ufs_support.html -

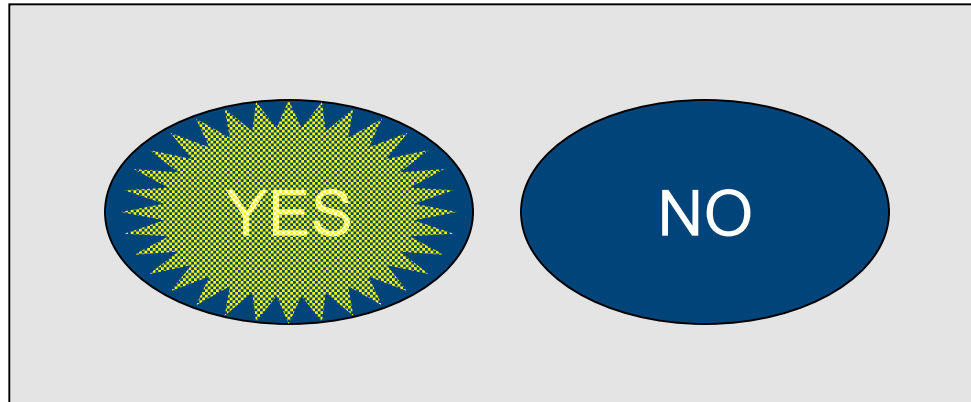
[Unicycle for Smalltalk](#)
... P980203.exe (140KB - 27/5/98) This fix provides s
paths" (ie \\ServerName\Dir\File.Ext) within the **UniCy**
www.unity-software.com/html/support/apkitpatches/a

[Goudurix ::](#)
... Diabolo • Phosphorescent Diabolo • **Repair** and Spare ... Stilts • St
• Fire Line ... French only) • **Instructions** Booklets • Videos ...
www.goudurix.com/produits_detail.php?ID=105&Ref=44 - 49k - Ca
- Similar pages

[Unicyclist Community Forums - KH 20 / KH 24 hub moving?](#)
... You've got three choices 1) Contact **Unicycle**.com and find out if they will replace
or **repair** the wheel 2) Send the wheel to Darren Bedford to fix 3) Fix it ...

Internet

**Did anyone here wonder what
UniCycle for SmallTalk is?**



It's your business to know.™

UniCycle for SmallTalk (so now you know!)



Products > UniCycle for Smalltalk

UniCycle for Smalltalk

Features

Software Maintenance Keeping you Informed and in Control

Unity's UniCycle is the first complete, integrated software maintenance system for VisualAge Smalltalk (www.software.ibm.com/ad/smalltalk), combining the functions of problem management, software distribution and task-based workflow systems. With the capability to manage the ongoing maintenance of software that is ready for deployment or has already been deployed, UniCycle keeps developers and managers fully informed and in control of their VisualAge applications.

Sample of UniCycle for Smalltalk Interface

Unique Features and Benefits of UniCycle:

- Remote Problem Capture and Fix
- Problem Management
- Software Patch Distribution
- Full-phase Capabilities
- Integration with Configuration Management System
- Capture of Debugger Stack as Objects
- Task Viewer
- Configuration Editor
- Easier Packaging
- Capture of Windows Registry Information
- Compression of Screen Capture Information
- Selective Window Capture for Manual Task Creation
- Separately Configurable Storage Targets for Capture and Patch
- Custom Attributes Specification within a Packaged Application



Features

Downloads

Case Study

Pricing Information

Support

It's your business to know.™

For High Precision, no beats Specificity

A single five character query
“v234r” finds the same single
document on both Google and
Yahoo.

The image shows two overlapping browser windows. The top window is Google, with the address bar showing `http://www.google.com/search?sourceid=navclient&ie=UTF-8&q=v234r`. The search results for 'v234r' are displayed, with the first result being a PDF document titled 'Effect of Mutation and Phosphorylation of Type I Keratins on Their Caspase-mediated Degradation -- Ku and Omary 276...'. The bottom window is Yahoo!, with the address bar showing `http://search.yahoo.com/search?p=v234r&csz=8&ei=UTF-8&web-t&cop=mss&tab=&toggle=1`. The search results for 'v234r' are also displayed, showing the same first result as Google. A yellow callout box points to the search bar in the Yahoo! window, containing the text 'A single five character query "v234r" finds the same single document on both Google and Yahoo.'

Address `http://www.google.com/search?sourceid=navclient&ie=UTF-8&q=v234r`

Google v234r

Web Images Groups News Froogle Desktop mo

Address `http://search.yahoo.com/search?p=v234r&csz=8&ei=UTF-8&web-t&cop=mss&tab=&toggle=1`

Google v234r

Yahoo! My Yahoo! Mail Welcome, Guest [Sign In] Search Home Help

Web Images Directory Local **NEW!** News Products

YAHOO! search v234r Search

Search Results Results 1 - 1 of about 1 for v234r - 0.17 sec. (About this page)

1. [Effect of Mutation and Phosphorylation of Type I Keratins on Their Caspase-mediated Degradation -- Ku and Omary 276...](#)
... arginine substitution at the X1 position of K18 (V234R) blocks K18 cleavage at Asp237 (Fig. 5B, lane 3 ... a (p29). Note that K18 V234R (lane 3) is not a caspase ...
www.jbc.org/cgi/content/full/276/29/26792 - More from this site

Web Images Directory Local **NEW!** News Products

Your Search: v234r Search

Help us improve your search experience. [Send us feedback.](#)
Create your own personal search experience with [My Yahoo! Search](#) [BETA]

Done

Internet

Keyword Matching



- We can achieve high precision when we provide very specific keywords to match
- Matching generic terms is not likely to produce sufficient precision for flagging sensitive content

False positives = Insufficient precision

It's your business to know.™

Document Clustering Techniques



- If you can process a representative set of your sensitive content then you can use document clustering techniques
- Examples...
 - HR documents
 - Source code for specific projects
 - Corporate financial documents
 - Marketing documents

It's your business to know.™

Document Clustering: “Documents like this”



- We can scan new content we see for overwhelming similarity to existing sensitive documents
- Documents can be viewed as similar based on statistical similarities or you can look for language similarities (or both)

It's your business to know.™

Text Analysis: Formatted Rich Text

A screenshot of a software application window displaying a document with formatted rich text. The window has a title bar with a search icon and a scroll bar on the right. The document content includes a header with "COMPANY NAME" and "PROCEDURES MANUAL", a table with four columns, a paragraph, a bulleted list, another table, another paragraph, and a second bulleted list. The status bar at the bottom shows "Page 3", "Sec 1", "3/5", and various function keys like "At", "Ln", "Col", "REC", "TRK", "EXT", "OVR".

COMPANY NAME PROCEDURES MANUAL

82203	Auditing	version 1	03/06/2001
-------	----------	-----------	------------

During the audit, the auditor will:

- make use of standard auditing techniques to collect objective information concerning the subject being audited
- not respond to rumour and hearsay
- avoid confrontational situations and arguments
- makes notes to aid the writing of the audit report
- collect documentary evidence of conformity or nonconformity
- note where current procedures could be improved
- keep the auditee informed as to the progress of the audit and any findings.

82204	Writing of the audit report	version 1	03/06/2001
-------	-----------------------------	-----------	------------

As soon as practical after the audit, the auditor will prepare an audit report which

- classifies the findings of the audit as follows:
 - * acceptable: satisfies the requirements of ISO 9001: 2000 and the company's own standards, procedures, manual, etc.
 - * major: fails to satisfy the requirements of ISO 9001: 2000
 - * minor: satisfies the requirements of ISO 9001: 2000 but fails to satisfy the company's own internal standards, procedures, manual, etc.

Page 3 Sec 1 3/5 At Ln Col REC TRK EXT OVR

Text Analysis:

Remove formatting, convert to unicode



```
82203 Auditing version 1 03/06/2001

During the audit, the auditor will:

make use of standard auditing techniques to collect
objective information concerning the subject being audited
not respond to rumour and hearsay
avoid confrontational situations and arguments
makes notes to aid the writing of the audit report
collect documentary evidence of conformity or nonconformity
note where current procedures could be improved
keep the auditee informed as to the progress of the audit
and any findings.

82204 Writing of the audit report version 1 03/06/2001

As soon as practical after the audit, the auditor will
prepare an audit report which

classifies the findings of the audit as follows:
acceptable: satisfies the requirements of ISO 9001: 2000
and the company's own standards, procedures, manual, etc.
major: fails to satisfy the requirements of ISO 9001: 2000
minor: satisfies the requirements of ISO 9001: 2000 but
fails to satisfy the company's own internal standards,
```

It's your business to know.™

Text Analysis: Tokenize, stem, normalize



```
82203 Auditing version 1 03/06/2001

During the audit, the auditor will:

make use of standard auditing techniques to collect
objective information concerning the subject being audited
not respond to rumour and hearsay
avoid confrontational situations and arguments
makes notes to aid the writing of the audit report
collect documentary evidence of conformity or nonconformity
note where current procedures could be improved
keep the auditee informed as to the progress of the audit
and any findings.

82204 Writing of the audit report version 1 03/06/2001

As soon as practical after the audit, the auditor will
prepare an audit report which

classifies the findings of the audit as follows:
acceptable: satisfies the requirements of ISO 9001: 2000
and the company's own standards, procedures, manual, etc.
major: fails to satisfy the requirements of ISO 9001: 2000
minor: satisfies the requirements of ISO 9001: 2000 but
fails to satisfy the company's own internal standards,
```

its your business to know.™

Document Representation:

What values to use for terms



- **tf (term frequency)** - Count of times term occurs in document.
 - The more times a term t occurs in document d the more likely it is that t is relevant to the document.
 - Used alone, favors common words, long documents.
- **df document frequency**
 - The more a term t occurs throughout all documents, the more poorly t discriminates between documents
- **tf-idf term frequency * inverse document frequency**
 - High value indicates that the word occurs more often in this document than average.

It's your business to know.™

Document Matching Based on Fingerprints



- Documents that are derived from the same source will likely have common fingerprints
- Documents that share common boilerplate information shared by many documents will not
- Documents talking about the same thing but not derived from the same source will probably not share common fingerprints

It's your business to know.™

Text Analysis:

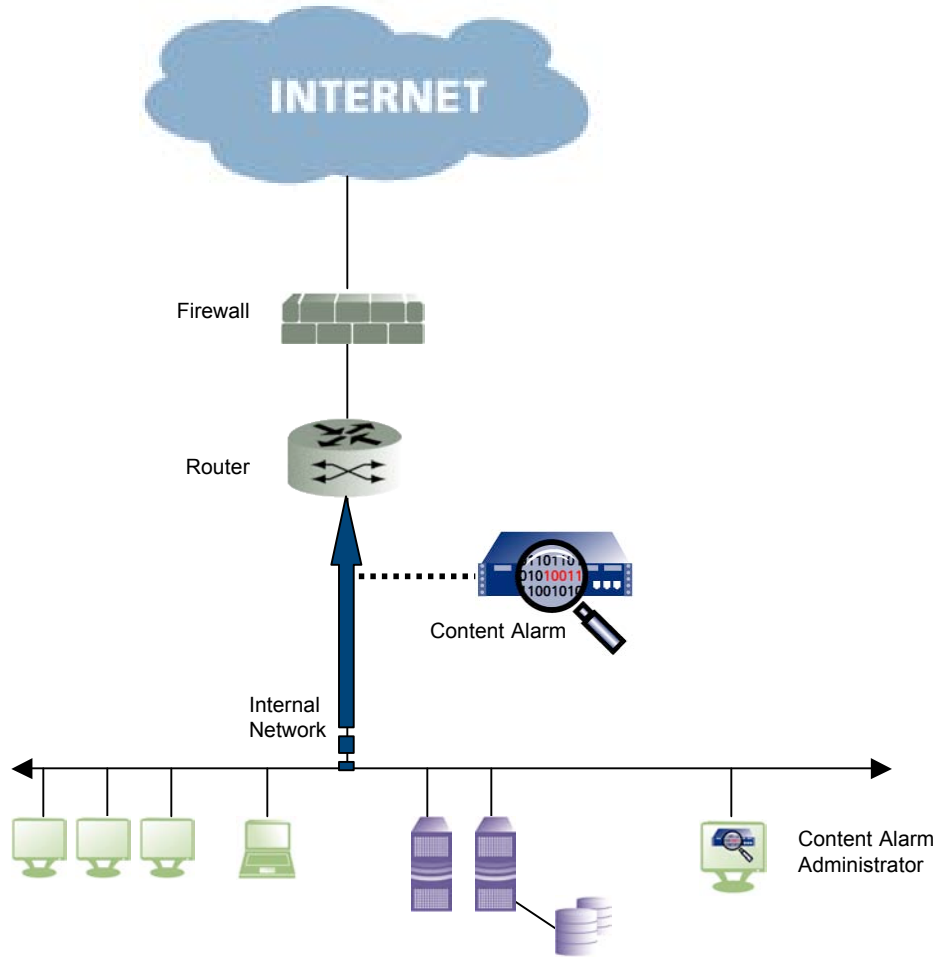
Generate collection of fingerprints



- Fingerprint 1 @1: 9aee9ea8d640be7f6df0766b02411e93
 - findings audit follows acceptable satisfies requirements ISO 9001 2000
- Fingerprint 2 @21: 4c7c05b25170b410398bffa26c7faa47
 - audit report version 2001 standard objective information concerning
- Fingerprint 3 @35: 8bfc73fa7a07aa92f091fce0b2272087
 - 82203 Auditing 2001 During audit auditor collect objective information
- ...

It's your business to know.™

Content Alarm Approach



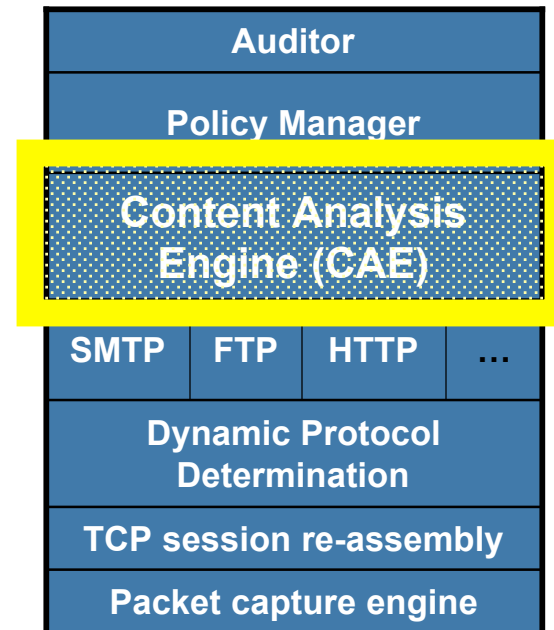
Auditor			
Policy Manager			
Content Analysis Engine (CAE)			
SMTP	FTP	HTTP	...
Dynamic Protocol Determination			
TCP session re-assembly			
Packet capture engine			

It's your business to know.™

Content Alarm Approach

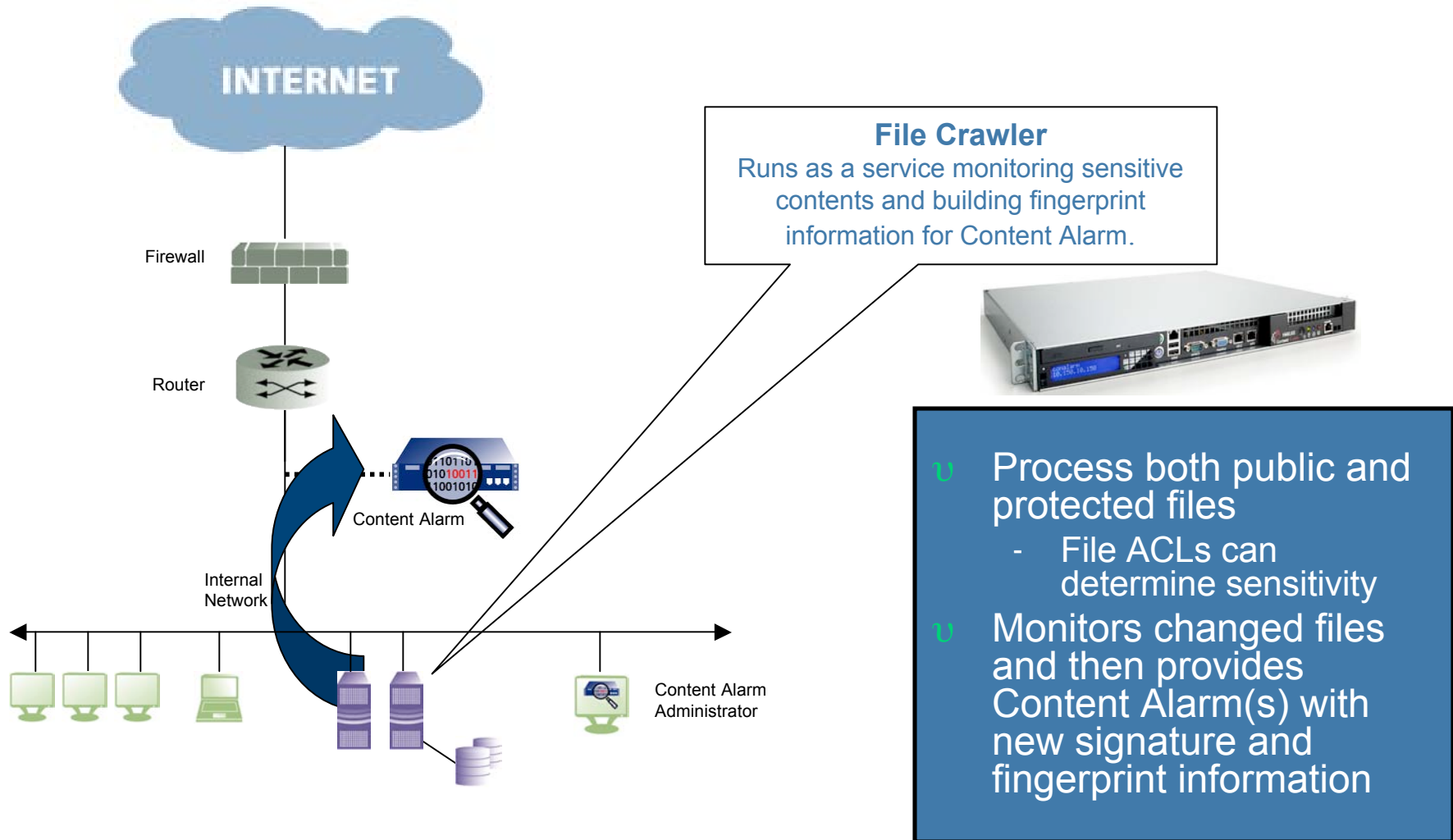


- Content Analysis Engine
 - Multiple analyzers
 - Keyword/regex analyzer
 - File signature analyzer
 - Meta data analyzer
 - Fingerprint analyzer
 - Database data analyzer*



* Version 2 feature

Content Alarm Architecture



Conclusion



- Think of search and knowledge discovery tools as part of your security toolkit!



It's your business to know.™

Contact Information



Tablus, Inc.
155 Bovet Road Suite 610
San Mateo, CA 94402

www.tablus.com

jim.nisbet@tablus.com

It's your business to know.™