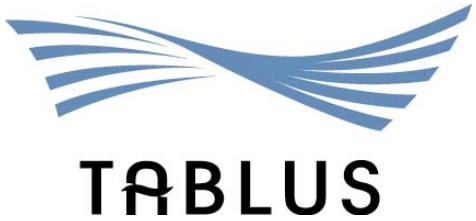


The Security Role for Content Analysis

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?

TABLUS

- 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?

TABLUS

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

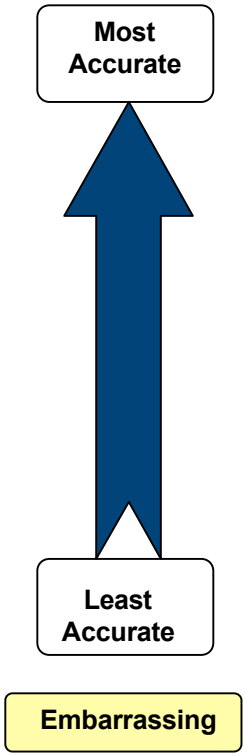
The challenge is to be able to accurately classify data as sensitive

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…



- The rights metadata tells us that it is (DRM/XRML)
- String/pattern matching
  - “confidential”, “do not distribute”, “\d{3}\-\d{2}\-\d{4}”
- Sensitive destination (domain, email id, etc.)
- Specific file extensions (C, CPP, XLS)
- Sensitive if file size exceeds specified size

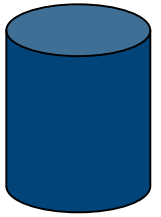
It’s your business to know.™

Types of Machine Classification Techniques

TABLUS

- 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)



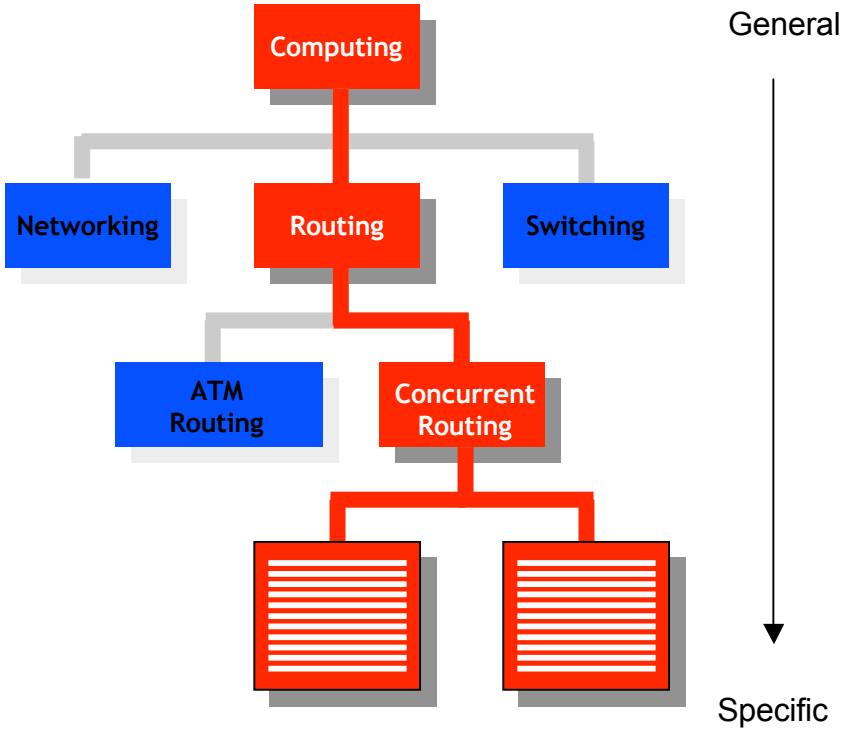
Training Set

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”

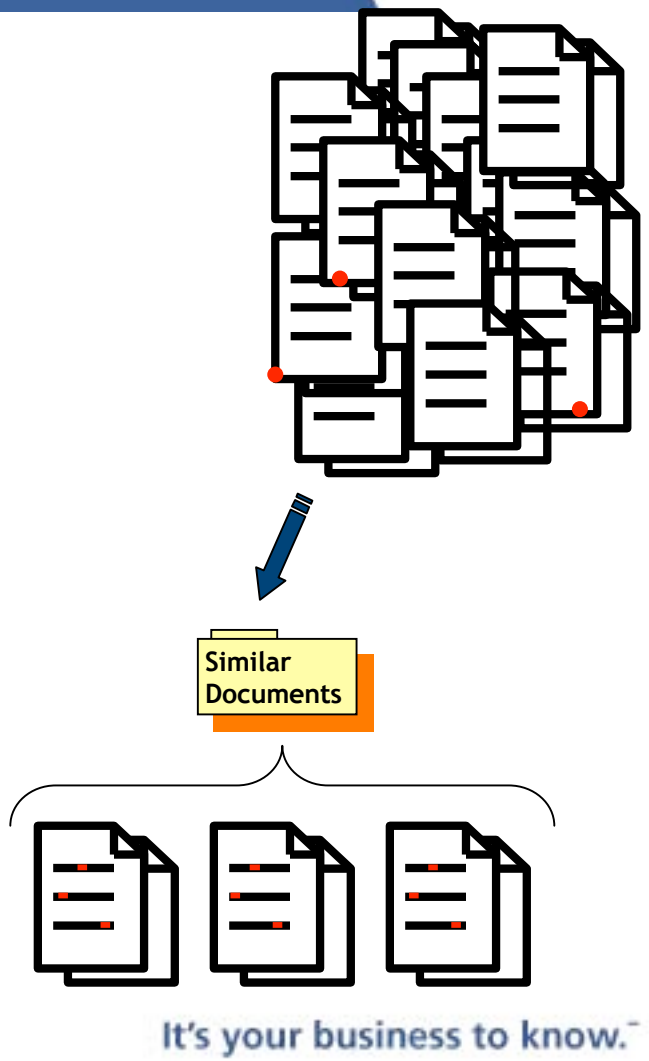


It’s your business to know.™

Document Clustering

TABLUS

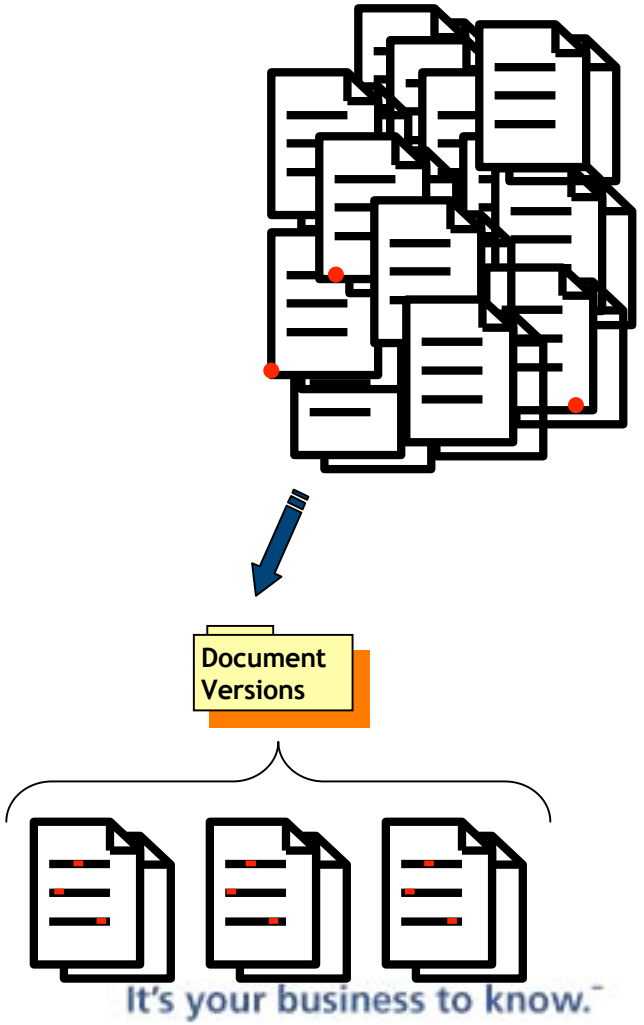
- Documents with similar characteristics grouped together
  - Based on language
  - Based on statistical information
  - Based on meta data



It's your business to know.™

Derivative Work Analysis

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



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

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

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

TABLUS

- 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?

TABLUS

ASSET CLASSIFICATION

It's your business to know.™

ISO 17799: Code of Practice for Information Security Management

TABLUS

- 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

TABLUS

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

TABLUS

- 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"?

TABLUS

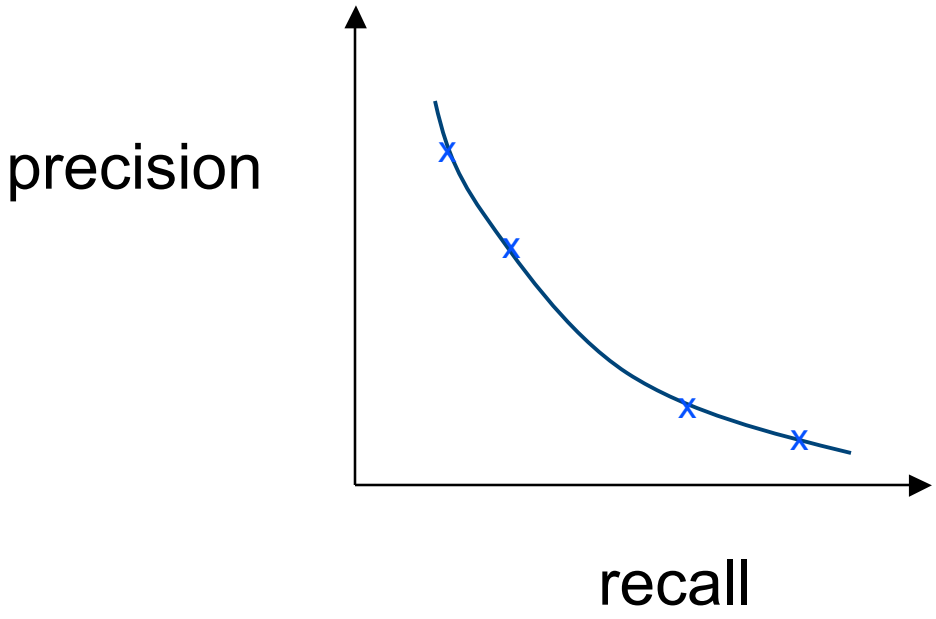
- Finding all occurrences is not too helpful
- Ranking the relevance of the results is key

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

Precision/Recall Curves

TABLUS

- 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

TABLUS

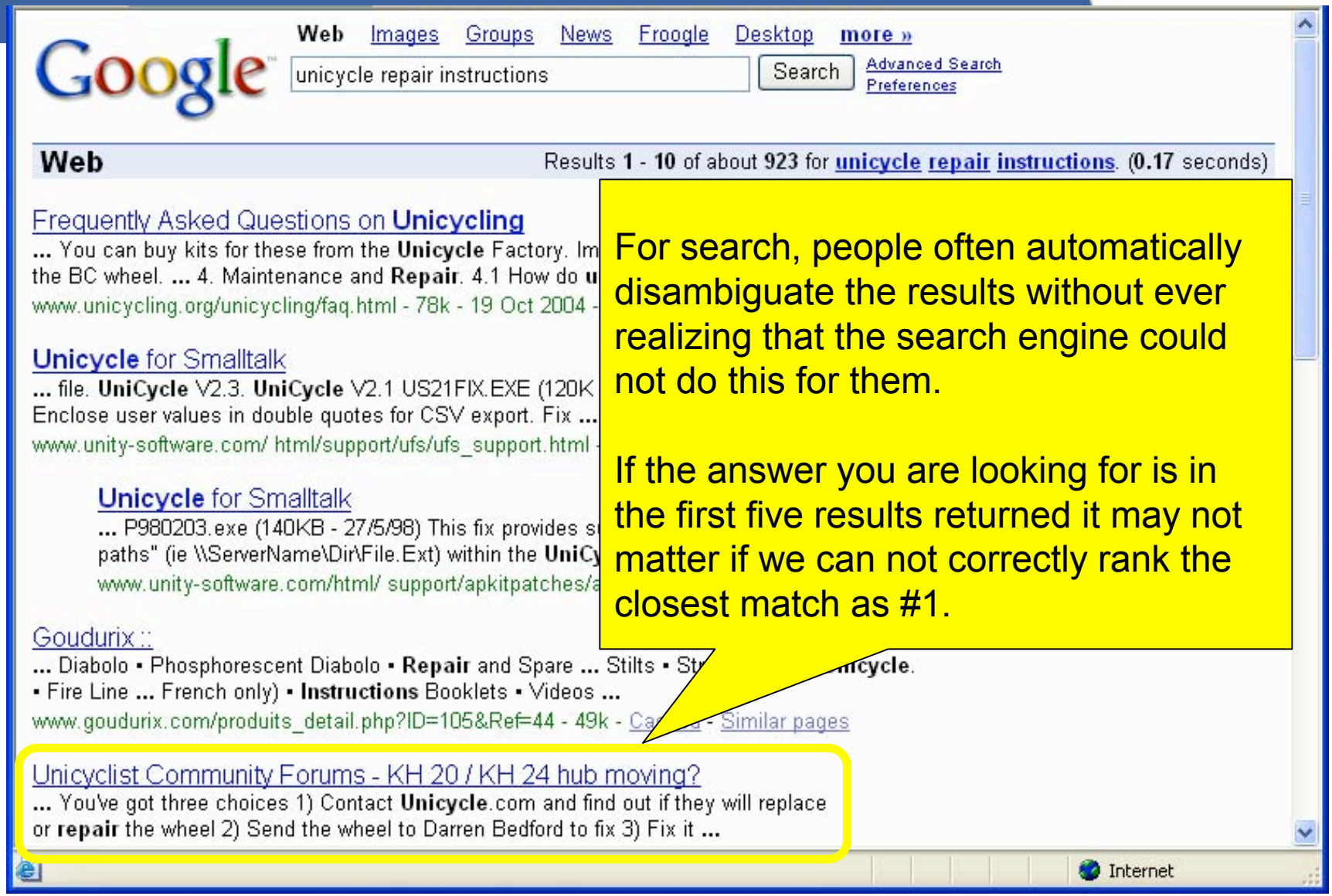
- Text REtrieval Conference/Competition
  - Run by NIST (National Institute of Standards & Technology)
- Collection: 3 Gigabytes, >1 Million Docs
  - Newswire & full text news (AP, WSJ, Ziff)
  - Government documents (federal register)
- Queries + Relevance Judgments
  - Queries devised and judged by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

An Example of Poor Precision...

TABLUS

Precision and Ambiguity

TABLUS

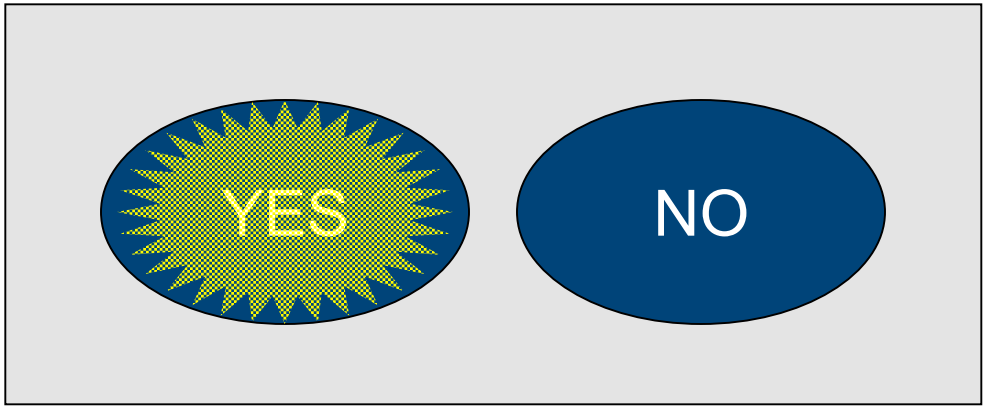


For search, people often automatically disambiguate the results without ever realizing that the search engine could not do this for them.

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.

Did anyone here wonder what UniCycle for SmallTalk is?

TABLUS



It's your business to know.™

UniCycle for SmallTalk (so now you know!)

TABLUS

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:

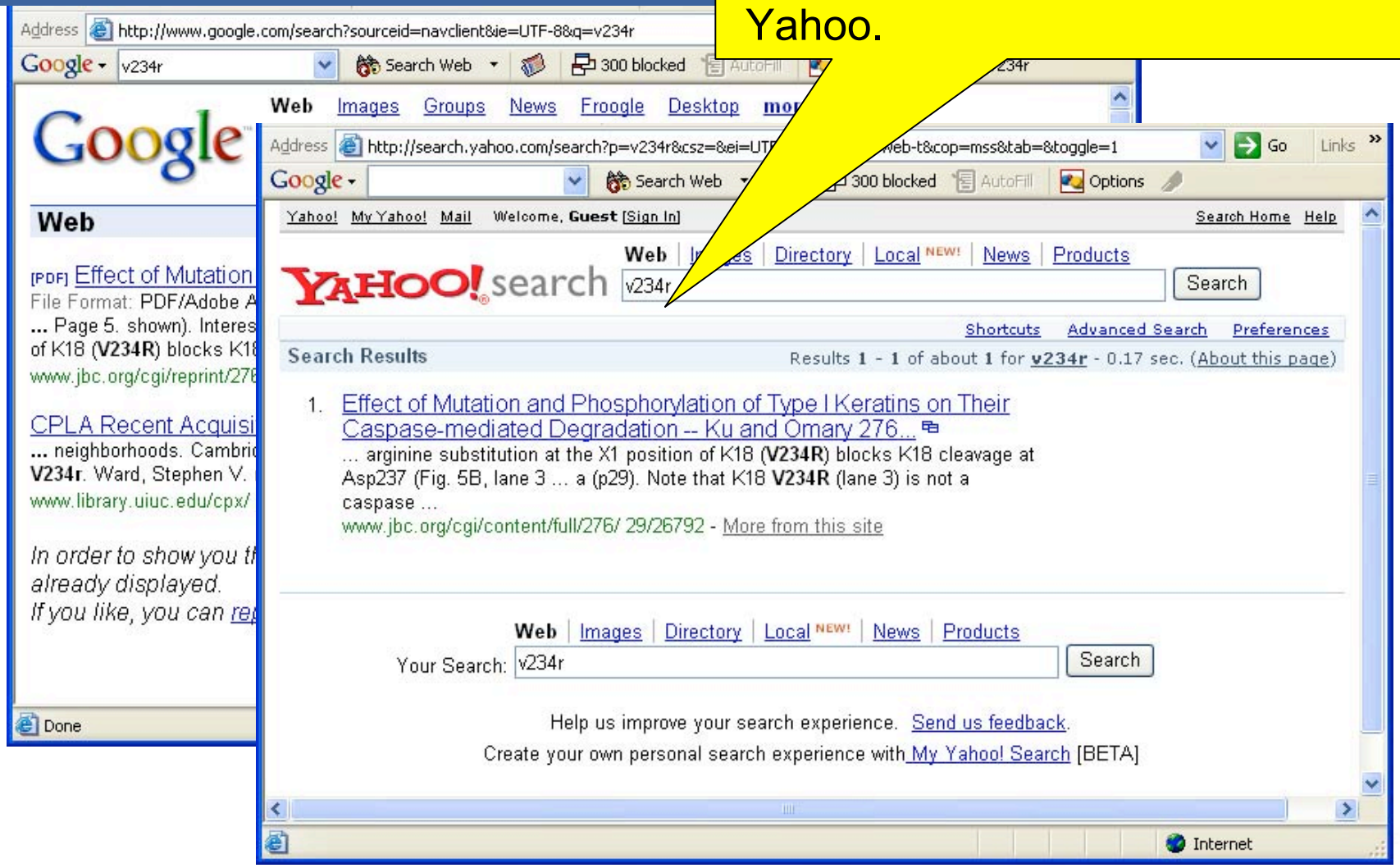
- Remote Problem Capture and Fix
- Problem Management
- Software Patch Distribution
- Full-phase Capabilities
- Integration with Configuration Management System
- Capture of Debugger Stack as Objects
- Task Viewer
- Configuration Editor
- Easier Packaging
- Capture of Windows Registry Information
- Compression 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, nc
beats Specificity

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



Keyword Matching

TABLUS

- 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

TABLUS

- 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"

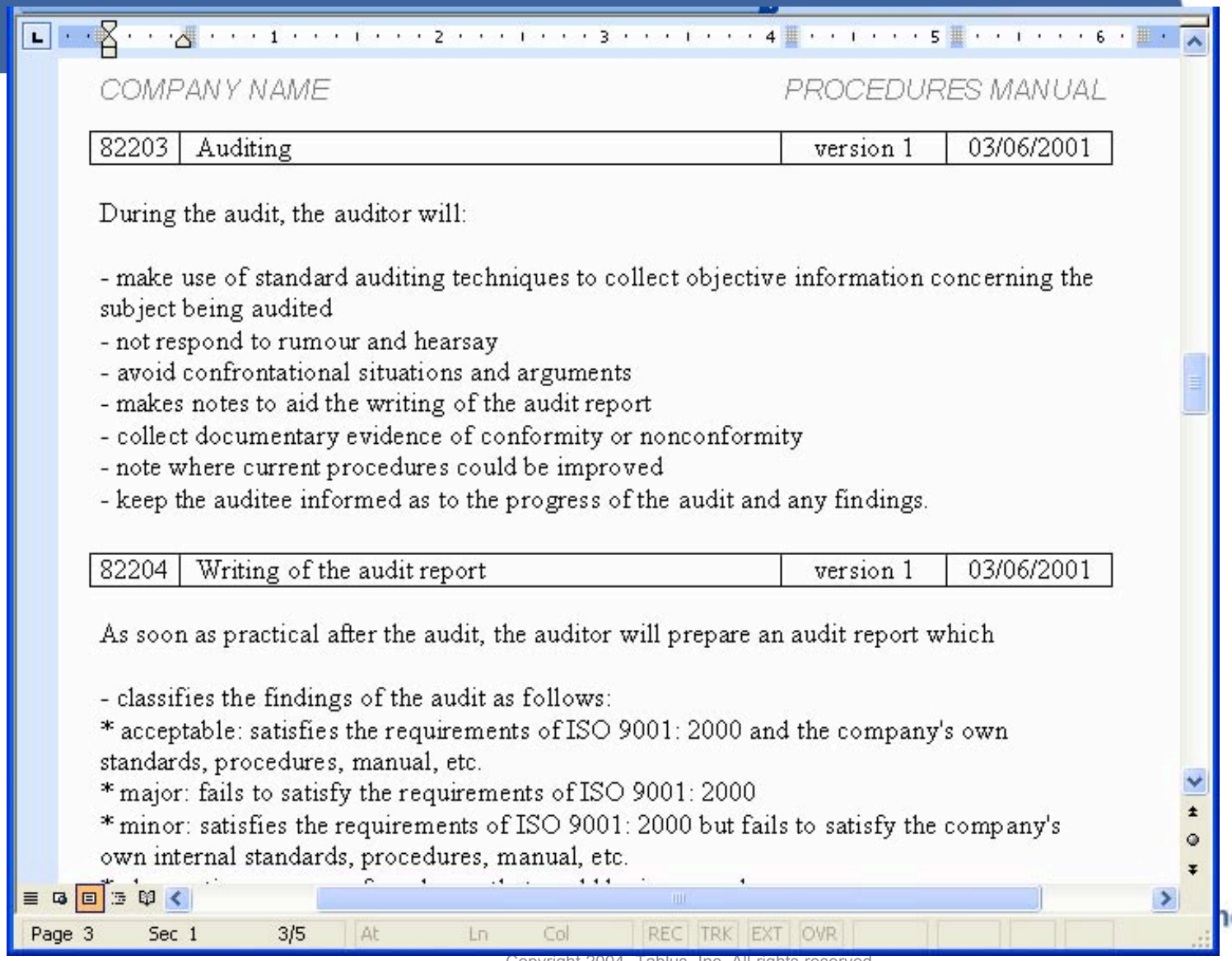
TABLUS

- 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



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.

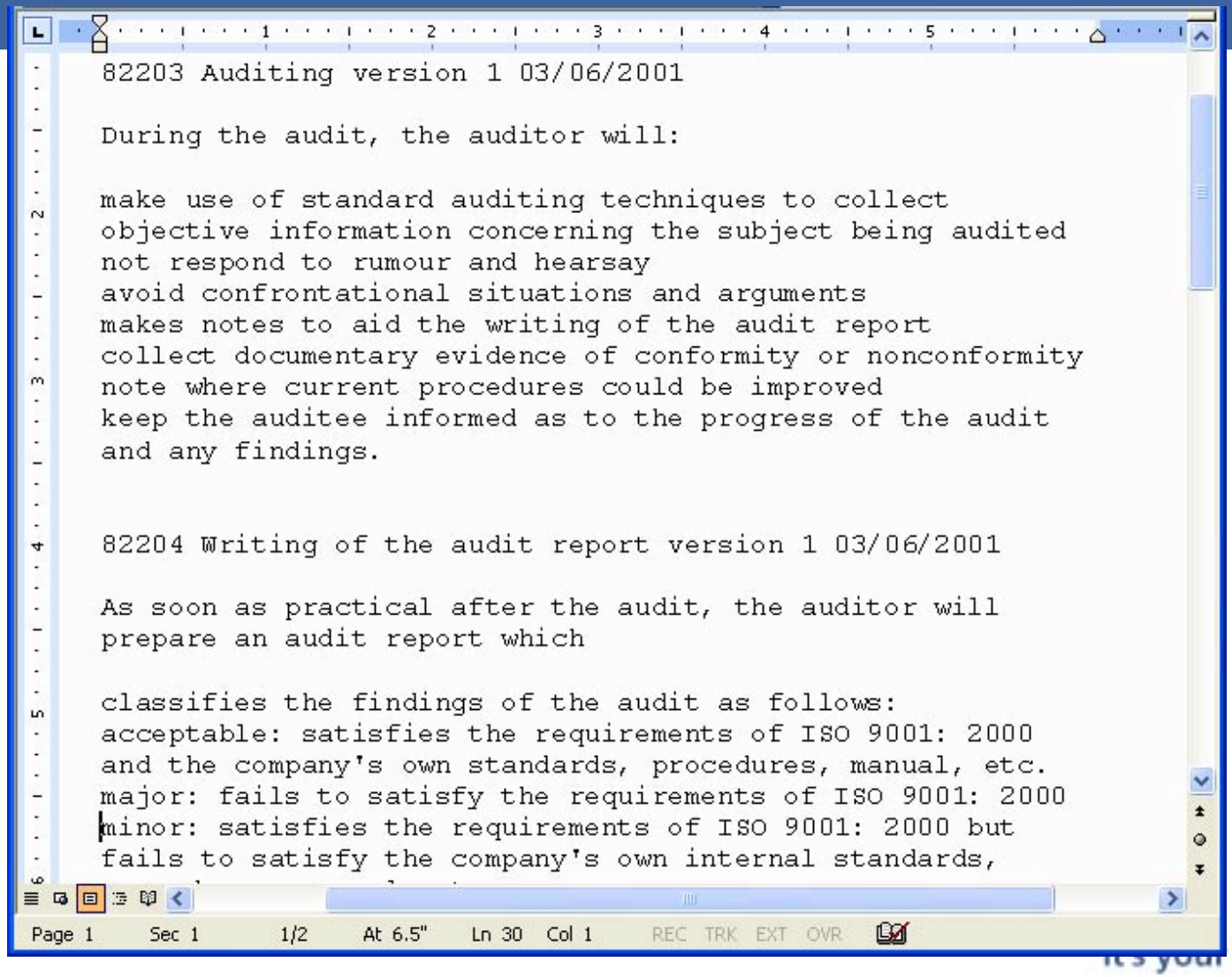
Page 3 Sec 1 3/5 At Ln Col REC TRK EXT OVR

TABLUS

ness to know.

Text Analysis:

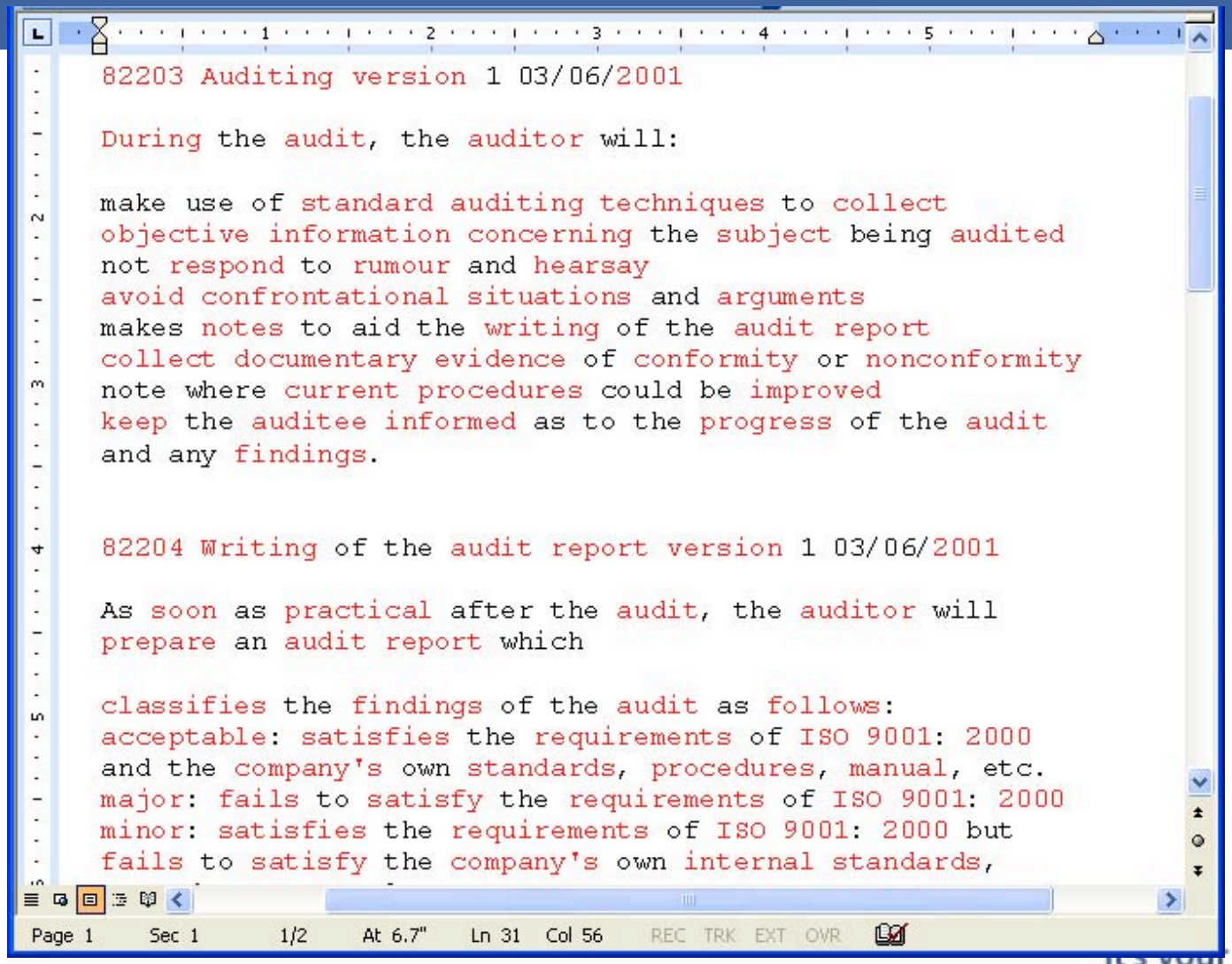
Remove formatting, convert to unicode



TABLUS

It's your business to know.

Text Analysis:
Tokenize, stem, normalize



TABLUS

It's 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.

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

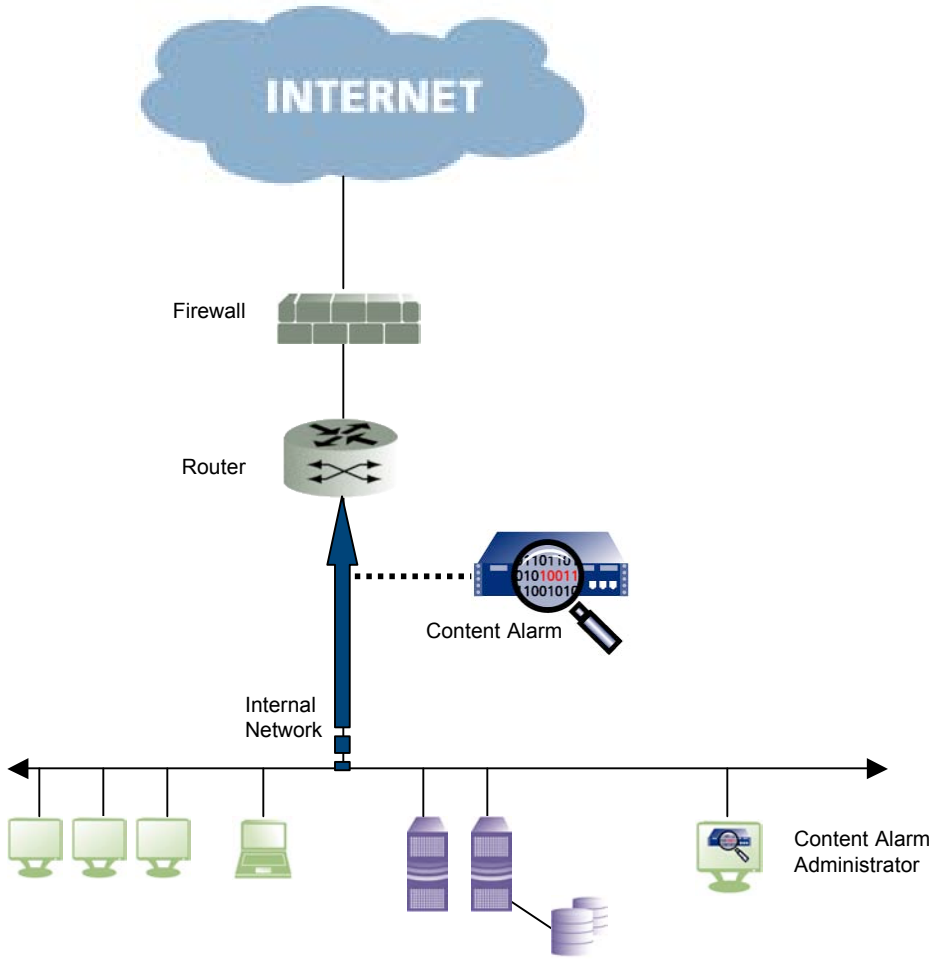
It's your business to know.™

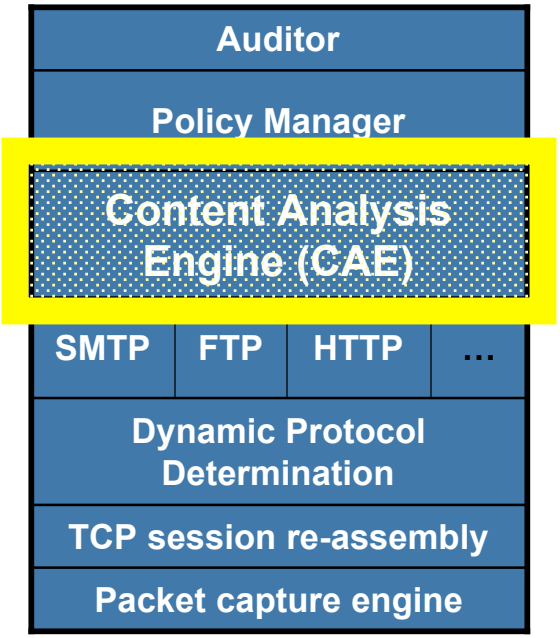
Text Analysis:
Generate collection of fingerprints

- 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

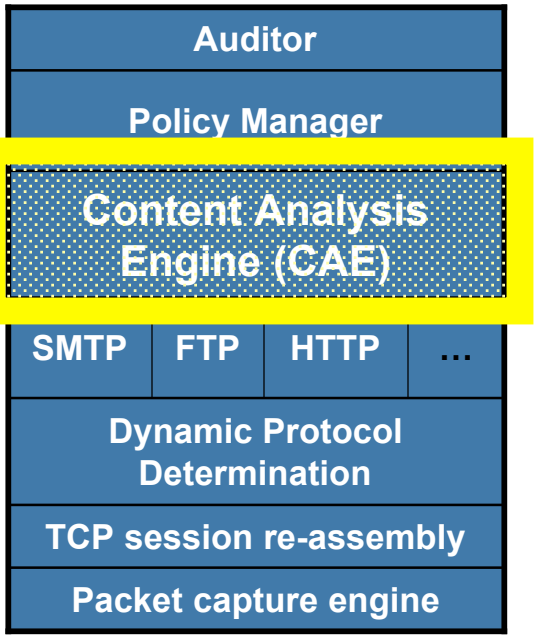




It's your business to know.™

Content Alarm Approach

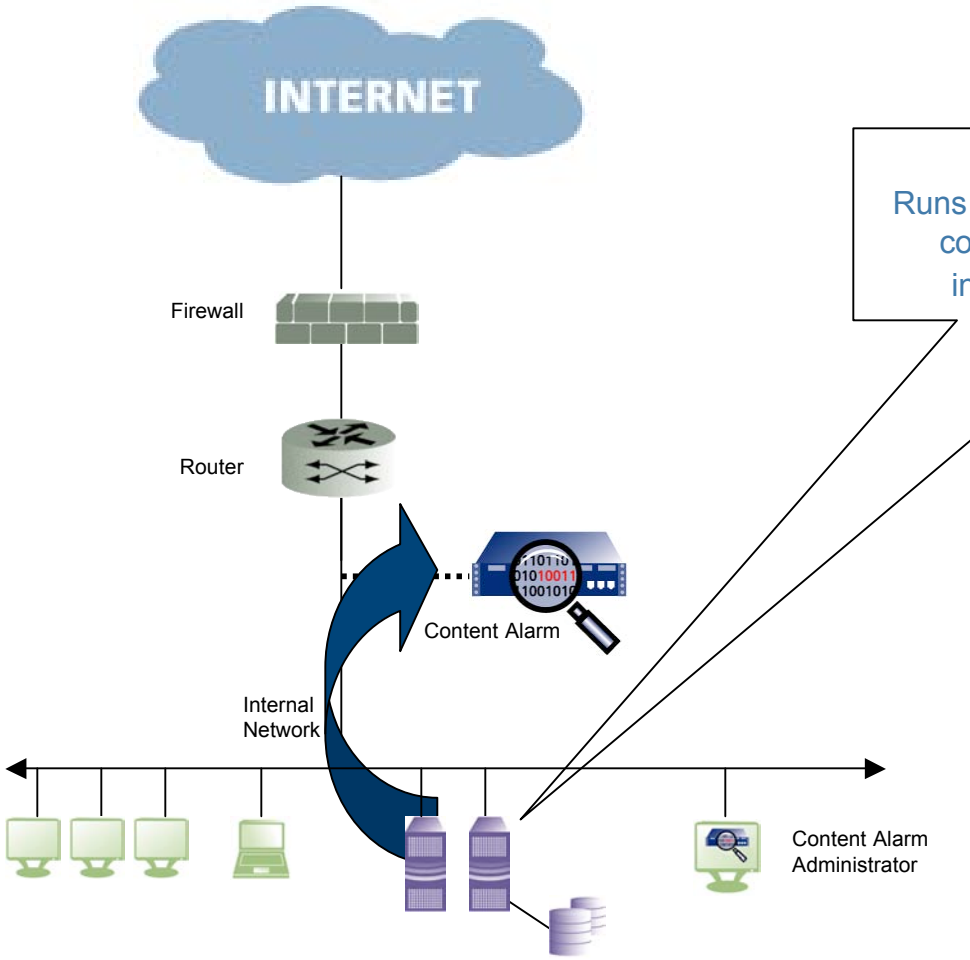
- Content Analysis Engine
  - Multiple analyzers
    - Keyword/regex analyzer
    - File signature analyzer
    - Meta data analyzer
    - Fingerprint analyzer
    - Database data analyzer*



* Version 2 feature

It's your business to know.™

Content Alarm Architecture



File Crawler
Runs as a service monitoring sensitive contents and building fingerprint information for Content Alarm.

- Process both public and protected files
  - File ACLs can determine sensitivity
- Monitors changed files and then provides Content Alarm(s) with new signature and fingerprint information

Conclusion

TABLUS

- 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.™