From Search Engines to Web Mining

“Web Search Engines, Spiders, Portals, Web APIs, and Web Mining: From the Surface Web and Deep Web, to the Multilingual Web and the Dark Web”

Hsinchun Chen, University of Arizona
Outline

• Google Anatomy and Google Story
• Inside Internet Search Engines (Excite Story)
• Vertical and Multilingual Portals: HelpfulMed and CMedPort
• Web Mining: Using Google, EBay, and Amazon APIs
• The Dark Web

The Google Story, by Vise and Malseed, 2005
Google Architecture

• Most Google is implemented in C or C++ and can run on Solaris or Linux
• URL Server, Crawler, URL Resolver
• Store Server, Repository
• Anchors, Indexer, Barrels, Lexicon, Sorter, Links, Doc Index
• Searcher, PageRank
• (See diagram)
PageRank

- $PR(A) = (1-d) + d \left( \frac{PR(T1)}{C(T1)} + \frac{PR(T2)}{C(T2)} + \ldots + \frac{PR(Tn)}{C(Tn)} \right)$

- Page A has T1…Tn pages which point to A.

- $d$ is a damping factor of $[0..1]$; often set as 0.85

- $C(T1)$ is the number of links going out of page T1.
Indexing

- Repository: Contains the full html page.
- Document Index: Keeps information about each document. Fixed with ISAM index, ordered by docID.
- Hit Lists: Corresponds to a list of occurrences of a particular word in a particular document including position, font, and capitalization information.
- Inverted Index: For every valid wordID, the lexicon contains a pointer into the barrel that wordID falls into. It points to a doclist of docID’s together with their corresponding Hit Lists.
Crawling

- Google uses a fast distributed crawling system.
- URL server and crawlers are implemented in Python.
- Each crawler keeps about 300 connections open at once.
- The system can crawl over 100 web pages (600K) per second using four crawlers.
- Follow “robots exclusion protocol” but not text warning.
Searching

• Ranking: A combination of PageRank and IR Score
• IR Score is determined as the dot product of the vector of count weights with the dot vector of type-weights (e.g., title, anchor, URL, plain text, etc.).
• User feedback to adjust the ranking function.
Storage Performance

• 24M fetched web pages
• Size of fetched pages: 147.8 GBs
• Compressed repository: 53.5 GBs
• Full inverted index: 37.2 GBs
• Total indexes (without pages): 55.2 GBs
Acknowledgements

• Hector Garcia-Molina, Jeff Ullman, Terry Winograd
• Stanford Digital Library Project (InfoBus/WebBase)
• Other DLI-1 projects: Berkeley, UCSB, UIUC, Michigan, and CMU
Google Story

• “They run the largest computer system in the world [more than 100,000 PCs].” John Hennessy, President, Stanford, Google Board Member

• PageRank technology
Google Story: VCs

• August 1998, met Andy Bechtolsheim, computer whiz and successfully angel; invested $100,000; Raised $1M from family and friends.
• “The right money from the right people led to the right contacts that could make or break a technology business.” ➔ The Stanford, Sand Hill Road contacts…
• John Doerr of Kleiner Perkins (Compaq, Sun, Amazon, etc.): $12.5M
• Miochael Moritz of Sequoia Capital (Yahoo): $12.5M
• Eric Schmidt as CEO (Ph.D. CS Berkeley, PARC, Bell Labs, Sun CEO)
Google Story: Ads

• “Banners are not working and click-through rates are falling. I think highly targeted focused ads are the answer.” – Brin ➔ “Narrowcast”

• Overture Inc ➔ GoTo’s money-making ads model

• Ads keyword auctioning system, e.g., “mesothelioma,” $30 per click.

• Network of affiliates that feature Google search on their sites.

• $440M in sales and $100M in profits in 2002.
Google Story: Culture

• 20% rule: Employees work on whatever projects interested them
• Hiring practice: flat organization, technical interviews
• IPO auction on Wall Street, “An Owners Manual for Google Shareholders”
• The only Chef job with stock options! (Executive chef Charlie Ayers)
• Gmail, Google Desktop Search, Google Scholar
• Google vs. Microsoft (FireFox)
Google Story: China

• Dr. Kia-Fu Lee, CMU Ph.D., founded Microsoft Research Asia in 1998; Google VP (President of Google China), 2006 ; Dr. Lee-Feng Chien, Google China Director

• Yahoo invested $1B in Alibaba (China e-commerce company)

• Baidu.com (#1 China SE) IPO in Wall Street, August 2005; stock soared from $27 to $122
Google Story: Summary

• Best VCs
• Best engineering
• Best engineers
• Best business model (ads)
• Best timing
• …so far
Beyond Google…

• Innovative use of new technologies…
• WEB 2.0, YouTube, MySpace…
• Build it and they will come…
• Build it large but cheap…
• IPO vs. M&A…
• Team work…
• Creativity…
• Taking risk…
Inside Internet Search Engines: Fundamentals

Jan Pedersen and William Chang

Excite

ACM SIGIR’99 Tutorial
Outline

- Basic Architectures
  - Search
  - Directory
- Term definitions:
  - Spidering, indexing etc.
  - Business model
Basic Architectures: Search

- Web
- Spider
- Log
- Index
- SE
- Spam
- Freshness
- 24x7
- 800M pages?
- 20M queries/day
- Browser
- Quality results

Quality results
Spidering

- Web HTML data
  - Hyperlinked
  - Directed, disconnected graph
  - Dynamic and static data
  - Estimated 800M indexible pages

- Freshness
  - How often are pages revisited?
Indexing

Size
- from 50 to 150M urls
- 50 to 100% indexing overhead
- 200 to 400GB indices

Representation
- Fields, meta-tags and content
- NLP: stemming?
Search

- Augmented Vector-space
  - Ranked results with Boolean filtering
- Quality-based reranking
  - Based on hyperlink data
  - or user behavior
- Spam
  - Manipulation of content to improve placement
1. **Information Retrieval at CNIDR**
   CNIDR?resources for information discovery and retrieval. U.S. Patents, Retrieval. Education Information. Service Information. CNIDR?Team. History...
   URL: [www.cnidr.org/infoir.html](http://www.cnidr.org/infoir.html)
   Translate  
   Related pages

2. **NIS Electronic Mathematical Information Retrieval (NIS-EMIR)**
   NIS-EMIR service. English version of our server. 
   URL: [www.ras.ruNIS/nis-emir-in.html](http://www.ras.ruNIS/nis-emir-in.html)
   More pages from this site  
   Related pages

3. **Information Retrieval and Analysis Group**
   The Information Retrieval and Analysis Group. Welcome to the home page for the Information Retrieval and Analysis Group at the T.J. Watson Research ...
Queries

- Short expressions of information need
  - 2.3 words on average
  - Relevance overload is a key issue
    - Users typically only view top results

- Search is a high volume business
  - Yahoo! 50M queries/day
  - Excite 30M queries/day
  - Infoseek 15M queries/day
Directory

- Manual categorization and rating
  - Labor intensive
    - 20 to 50 editors
  - High quality, but low coverage
    - 200-500K urls

- Browsable ontology

- Open Directory is a distributed solution
**Business Model**

- **Advertising**
  - Highly targeted, based on query
  - Keyword selling; Between $3 to $25 CPM
- **Cost per query is critical**
  - Between $.5 and $1.0 per thousand
- **Distribution**
  - Many portals outsource search
Web Resources

- Search Engine Watch
  - www.searchenginewatch.com
- “Analysis of a Very Large Alta Vista Query Log”; Silverstein et al.
  - SRC Tech note 1998-014
  - www.research.digital.com/SRC
Web Resources

- “The Anatomy of a Large-Scale Hypertextual Web Search Engine”; Brin and Page
  - google.stanford.edu/long321.htm
- WWW conferences
  - www8.org
Inside Internet Search Engines: Spidering and Indexing

Jan Pedersen
and
William Chang
Basic Architectures: Search

20M queries/day

Web

Spam

Freshness

800M pages?

Index

Log

SE

SE

SE

Browser

Quality results

24x7

Spider
Basic Algorithm

1. Pick Url from pending queue and fetch
2. Parse document and extract href’s
3. Place unvisited Url’s on pending queue
4. Index document
5. Goto (1)
**Issues**

- Queue maintenance determines behavior
  - Depth vs breadth
  - Spidering can be distributed
    - but queues must be shared
- Urls must be revisited
  - Status tracked in a Database
  - Revisit rate determines *freshness*
  - SE’s typically revisit every url monthly
Deduping

- Many urls point to the same pages
  - DNS aliasing
- Many pages are identical
  - Site mirroring
- How big is my index, really?
Smart Spidering

- Revisit rate based on modification history
  - Rapidly changing documents visited more often
  - Revisit queues divided by priority
- Acceptance criteria based on quality
  - Only index quality documents
  - Determined algorithmically
Spider Equilibrium

- Urls queues do not increase in size
  - New documents are discovered and indexed
  - Spider keeps up with desired revisit rate
  - Index drifts upward in size

- At equilibrium index is *Everyday Fresh*
  - As if every page were revisited every day
  - Requires 10% daily revisit rates, on average
Computational Constraints

- Equilibrium requires increasing resources
  - Yet total disk space is a system constraint
- Strategies for dealing with space constraints
  - Simple refresh: only revisit known urls
  - Prune urls via stricter acceptance criteria
  - Buy more disk
Special Collections

- Newswire
- Newsgroups
  - Specialized services (Deja)
- Information extraction
  - Shopping catalog
  - Events; recipes, etc.
The Hidden Web

- Non-indexible content
  - Behind passwords, firewalls
  - Dynamic content
  - Often searchable through local interface

- Network of distributed search resources
  - How to access?
  - Ask Jeeves!
Spam

- Manipulation of content to affect ranking
  - Bogus meta tags
  - Hidden text
  - Jump pages tuned for each search engine
- Add Url is a spammer’s tool
  - 99% of submissions are spam
- It’s an arms race
Representation

- For precision, indices must support phrases
  - Phrases make best use of short queries
  - The web is precision biased
- Document location also important
  - Title vs summary vs body
- Meta tags offer a special challenge
  - To index or not?
The Role of NLP

- Many Search Engines do not stem
  - Precision bias suggests conservative term treatment
- What about non-English documents
  - N-grams are popular for Chinese
  - Language ID anyone?
Inside Internet Search Engines: Search

Jan Pedersen
and
William Chang
Basic Architectures: Search

- Web
- Spider
- Spam
- Freshness
- Index
- Log
- SE
- 24x7
- 20M queries/day
- Quality results
- Browser
- 800M pages?
Query Language

- Augmented Vector space
  - Relevance scored results
  - Tf, idf weighting
  - Boolean constraints: +, -
  - Phrases: “”
- Fields:
  - e.g. title:
Does Word Order Matter?

- Try “information retrieval” versus “retrieval information”
  - Do you get the same results?

- The query parser
  - Interprets query syntax: +, -, “”
    - Rarely used
  - General query from free text
    - Critical for precision
Precision Enhancement

- Phrase induction
  - All terms, the closer the better
- URL and Title matching
- Site clustering
  - Group URLs from same site
- Quality-based reranking
Link Analysis

- Authors vote via links
  - Pages with higher inlink are higher quality
- Not all links are equal
  - Links from higher quality sites are better
  - Links in context are better
- Resistant to Spam
  - Only cross-site links considered
Page Rank (Page’98)

- Limiting distribution of a random walk
  - Jump to a random page with Prob. $\varepsilon$
  - Follow a link with Prob. $1 - \varepsilon$

- Probability of landing at a page D:
  - $\frac{\varepsilon}{T} + \sum P(C)/L(C)$
  - Sum over pages leading to D
  - $L(C) = $ number of links on page D
HITS (Kleinbergy’98)

- Hubs: pages that point to many good pages
- Authorities: pages pointed to by many good pages
- Operates over a vincity graph
  - pages relevant to a query
- Refined by the IBM Clever group
  - further contextualization
Hyperlink Vector Voting (Li’97)

- Index documents by in-link anchor texts
  - Follow links backward
  - Can be both precision and recall enhancing
    - The “evil empire”
- How to combine with standard ranking?
  - Relative weight is a tuning issue
Evaluation

- No industry standard benchmark
  - Evaluations are qualitative
  - Excessive claims abound
  - Press is not be discerning

- Shifting target
  - Indices change daily
  - Cross engine comparison elusive
Novel Search Engines

- Ask Jeeves
  - Question Answering
  - Directory for the Hidden Web
- Direct Hit
  - Direct popularity
  - Click stream mining
Summary

- Search Engines are surprisingly effective
  - Given short queries
  - Precision enhancing techniques are critical
- Centralized search is maximally efficient
  - but one can achieve a big index through layering
Inside Internet Search Engines: Business

William Chang
and
Jan Pedersen
Outline

- Business Evolution
  - From Search Engine to New Media Network

- Trends
  - Differentiation
  - Localization and Verticals

- The New Networks
  - Broadband
Search Engine Evolution

- Cataloguing the web
- Inclusion of verticals
- Acquisition of communities
  - Commercialization; localization
- The new networks
  - Keiretsu – linked by mutual obligation
  - Access
Cataloguing the web – human or spider?

- YAHOO! directory
- Infoseek Professional
  - quality content, $0.10/query = 20,000 users
- Web Search Engines
  - ....content, FREE = 50,000,000 users
- Sex and progress
- Community directory, community search
Inclusion of Verticals

- Content is king?
- Content or advertising?
- When you want content, they pay; when you need content, you pay
- Channels – pulling users to destinations through search
Acquisition of Communities

- Email, killer app of the internet
  - Mailing lists
- Usenet Newsgroups
- Bulletin boards
- Chat rooms
- Instant messaging
  - buddy lists, ICQ (I Seek You)
Community Commercialization

- Amazon
  - trusted communities to help people shop
- Ebay
  - collectors are early adopters (rec.collecting.*)
- B2B or C2C or B2C or C2B, who cares?
- ConsumerReview
- SiliconInvestor and YAHOO! Finance
  - Community and commerce are two sides of the same “utility” coin
Localization of Verticals

- Real-world portals
  - newspapers
- CitySearch, Zip2, Sidewalk, Digital Cities
  - whither local portals?
- Local queries
- Vertical comes first
- Our social fabric is interwoven from local and vertical interests
Differentiation?

- ABC, NBC, CBS – what’s the difference?
- Amusement park – YAHOO!
- TV – Excite
- Community center – Lycos
- Transportation – Infoseek
- Bus stops becoming bus terminal – Netscape
The New Networks

- A consumer revolution
  - The community makes the brand
  - Winning brands empower consumers, embrace the internet’s viral efficiency
- Media is at the core of brand marketing
- From portals to networks
  - navigation, advertising, commerce
The New Network

Ingredients:
- Search engine audience
- Ad agency
- Old media
- Verticals
- Bank
- Venture capital
- Access, technology, and services providers
Keiretsu

- SoftBank
  - YAHOO!, Ziff-Davis, NASDAQ?
- Kleiner Perkins
  - AOL, Concentric, Sun, Netscape, Intuit, Excite
- Microsoft
  - MSN, MSNBC, NBC, CNET, Snap, Xoom, GE
- AT&T
  - TCI, AtHome, Excite
Keiretsu

- CMGI
  - AltaVista, Compaq/DEC, Engage
- Lycos
  - WhoWhere, Tripod
- Disney
  - (ABC, ESPN), Infoseek (GO Network)
Access

- Broadband market
- Ubiquitous access or “convergence” of internet and telephony
- The other universal resources locator – the telephone number
- Wireless, wireless, wireless
HelpfulMED: Creating a Knowledge Portal for Medicine

Gondy Leroy and Hsinchun Chen
Heterogeneous Medical Literature Databases and the Internet

The Medical Information Gap

Medical Professionals & Users

Current Information Interfaces
Research Questions

• How can linguistic parsing and statistical analysis techniques help extract medical terminology and the relationships between terms?

• How can medical and general ontologies help improve extraction of medical terminology?

• How can linguistic parsing, statistical analysis, and ontologies be incorporated in customizable retrieval interfaces?
Previous Work:

Linguistic Parsing and Statistical Analysis
Noun compounds are widely used across sub-language domains to describe concepts concisely.

Unlike keyword searching, contextual information is available.

Relationship between a noun compound and the head noun is a strict conceptual specification.
- “breast” and “cancer” vs. “breast cancer”
- “treatment” and “cancer” vs. “treatment of cancer”

Proper nouns can be captured

(Anick and Vaithyanathan, 1997)
Appropriate level of analysis: Extraction of grammatically correct noun phrases from free text

Used in other domains, noun phrasing has been shown to improve the accuracy of information retrieval (Girardi, 1993; Devanbu et al., 1991; Doszkocs, 1983)

Cooper and Miller (‘98) used noun phrasing to map user queries to MeSH with good results
Arizona Noun Phraser

- NSF Digital Library Initiative I & II Research
- Developed to improve document representation and to allow users to enter queries in natural language
• **Tokenizer**
  – Takes raw text and generates word tokens (conforms to UPenn Treebank word tokenization rules)
  – Separates punctuation and symbols from text without affecting content

• **Part of Speech (POS) Tagger**
  – Based on the Brill Tagger
  – Two-pass parser, assigns parts of speech to each word
  – Uses both lexical and contextual disambiguation in POS assignment
  – Lexicons: Brown Corpus, Wall Street Journal, Specialist Lexicon

• **Phrase Generation**
  – Simple Finite State Automata (FSA) of noun phrasing rules
  – Breaks sentences and clauses into grammatically correct noun phrases
• Results of Testing (Tolle & Chen, 1999)

The Arizona Noun Phraser is better than or comparable to other techniques (MIT’s Chopper and LingSoft’s NPtool).

• Improvement with Specialist Lexicon

The addition of the Specialist Lexicon to the other non-medical lexicons slightly improved the Arizona Noun Phraser’s ability to properly identify medical terminology.
Creating Knowledge Sources: Concept Space (Automatic Thesaurus)

- **Statistical Analysis Techniques:**
  - Based on document term co-occurrence analysis, weights between concepts establish the strength of the association.
  - Four steps: Document Analysis, Concept Extraction, Phrase Analysis, Co-occurrence Analysis.

- **Systems:**
  - Other: Geographical Information Systems, Law Enforcement.

- **Results:**
  - Alleviate cognitive overload, improve search recall.
The computation generated Cancer Space, which consists of 1.3M cancer terms and 52.6M cancer relationships.

The approach: **Object-Oriented Hierarchical Automatic Yellowpage (OOHAY)** -- the reverse of YAHOO!

Prototype system available for web access at: ai20.bpa.arizona.edu/cgi-bin/cancerlit/cn

Experiments for 10M Medline abstracts and 50M Web pages under way.
The Arizona team used NCSA’s 128-processor Origin2000 for over 20,000 CPU-hours.

Cancer Map used 1M CancerLit abstracts to generate 21,000 cancer topics in a 5-layer hierarchy of 1,180 cancer maps.

The research is part of the Arizona OOHAY project funded by NSF Digital Library Initiative 2 program.

Techniques: computational linguistics and neural network text mining.
Medical Concept Mapping:

Incorporating Ontologies
(WordNet and UMLS)
Incorporating Knowledge Sources: WordNet Ontology

- Princeton, George A. Miller (psychology dept.)
- 95,600 different word forms, 57,000 nouns
- grouped in synsets, uses word senses
- used to extract textual contexts (Stairmand, 1997), text retrieval (Voorhees, 1998), information filtering (Mock & Vermuri, 1997)
- available online: http://www.cogsci.princeton.edu/~wn/
WordNet 1.6 overview for "head"

The noun "head" has 30 senses in WordNet.

1. head, caput -- (the upper or front part of the body in animals; contains the face and brains; "he stuck his head out the window")
2. head -- (a single domestic animal: "200 head of cattle")
3. mind, head, brain, psyche, nous -- (that which is responsible for one's thoughts and feelings; the seat of the faculty of reason; "his mind wandered"; "I couldn't get his words out of my head")
4. head, chief, top dog -- (a person who is in charge; "the head of the whole operation")
5. head -- (the front of a military formation or procession; "the head of the column advanced boldly"; "they were at the head of the attack")
6. head -- (the pressure exerted by a fluid; "a head of steam")
7. head -- (the top of something; "the head of the stairs", "the head of the page", "the head of the list")
8. fountainhead, headspring, head -- (the source of water from which a stream arises; "they tracked him back toward the head of the stream")
9. head, head word -- ((linguistics) the word in a constituent that plays the same grammatical role as the whole)
10. head -- (the tip of an abscess (where the pus accumulates))
11. head -- (the length or height based on the size of a human or animal head; "he is two heads taller than his little sister"; "his horse won by a head")
12. capitulum, head -- (a dense clusters of flowers or foliage: "a head of cauliflower"; "a head of
Incorporating Knowledge Sources: UMLS Ontology

- **Unified Medical Language System** (UMLS) by the National Library of Medicine (Alexa McCray)

- 1986 - 1988: defining the user needs and the different components


- 1992 - present: updating & expanding the components, development of applications

UMLS Metathesaurus (2000 edition)

- 730,000 concepts, 1.5 M concept names
- 60+ vocabulary sources integrated
- 15 different languages
- organization by concept, for each concept there are different string representations
UMLS Metathesaurus (2000 edition)
• **134 semantic types** and **54 semantic relations**

• metathesaurus concepts $\rightarrow$ semantic net

• relations between types, not between concepts

---


Semantic Type: Pharmacologic Substance  
(105,784 concepts)

- aspirin
- heroin
- diuretics
- .......

Semantic Type: Sign or Symptom  
(4,364 concepts)

- aphasia
- aspirin allergy
- headache
- .......

- treats
- is a
- treats

"Pharmacologic Substance treats"

Type: Pharmacologic Substance
Relation: treats

Related Types:
- Anatomical Abnormality
- Acquired Abnormality
- Congenital Abnormality
- Injury or Poisoning
- Pathologic Function
- Cell or Molecular Dysfunction
- Disease or Syndrome
- Mental or Behavioral Dysfunction
- Neoplastic Process
- Experimental Model of Disease
- Sign or Symptom
UMLS Specialist Lexicon (2000 edition)

- A general English lexicon that includes many biomedical terms
- 130,000+ entries
- each entry contains syntactic, morphological and orthographic information
- no different entries for homonyms
UMLS Knowledge Source Server

Specialist Lexical Record

{base=blood type
 entry=EO013518
 cat=noun
 variants=reg
 }

View "blood type" in relational format.
Ontology-Enhanced Concept Mapping: Design and Components

Natural Language Component

Synonyms provided by WordNet and UMLS Metathesaurus

Related Concepts provided by Concept Space and limited with Deep Semantic Parsing (based on UMLS Semantic Net)
**Synonyms**

- **WordNet**
  - Return synonyms if there is only one word sense for the term
  - E.g. “cancer” has 4 different senses, one of them is:
    - Cancer, Cancer the Crab, fourth sign of the Zodiac

- **UMLS Methathesaurus**
  - find the underlying concept of a term and retrieve all synonyms belonging to this concept
  - E.g. term = tumor ⇒ concept = neoplasm
    - **synonyms:**
      - Neoplasm of unspecified nature NOS | tumor <1> | Unspecified neoplasms | New growth | [M]Neoplasms NOS | Neoplasia | Tumour | Neoplastic growth | NG - Neoplastic growth | NG - New growth | 800 NEOPLASMS, NOS |

- **filtering of the synonyms (personalizable for each user):** filter out the terms
  - tumor <1> | [M]Neoplasms NOS | NG - Neoplastic growth | NG - New growth | 800 NEOPLASMS, NOS |
Related Concepts

- **Retrieve related concepts** for all search terms from Concept Space
- **Limit related concepts** based on Deep Semantic Parsing
  (by means of the UMLS Semantic Net)

**Deep Semantic Parsing - Algorithm**

- **Step 1:** establish the semantic context for each original query (find the semantic types and relations of the search terms)
- **Step 2:** for each related concept, find if it fits the established context
- **Step 3:** reorder the final list based on the weights of the terms (relevance weights from CancerSpace)
- **Step 4:** select the best terms (highest weights) from the reordered list
Are lymph nodes and stromal cells related to each other?

Concept Space Terms

(filtered by Semantic Net):
- bone marrow
- lymphatic metastasis
- lymph node metastases
Medical Concept Mapping:

User Validation
User Studies

• Study 1: Incorporating Synonyms
• Study 2: Incorporating Related Concepts

• Input:
  – 30 actual cancer related user-queries

• Input Method:
  – Original Queries
  – Cleaned Queries
  – Term Input

• Golden Standards:
  – by Medical Librarians
  – by Cancer Researchers

• Recall and Precision:
  – based on the Golden Standards
Example of a Query

• **Original Query:** “What causes fibroids and what would cause them to enlarge rapidly (patient asked Dr. B and she didn’t know)”

• **Cleaned Query:** “What causes fibroids and what would cause them to enlarge rapidly?”

• **Term input:** “fibroids”
<table>
<thead>
<tr>
<th></th>
<th>Medical Librarians</th>
<th>Cancer Researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. Terms per Query:</td>
<td>39</td>
<td>9</td>
</tr>
<tr>
<td>Min. Terms per Query:</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Average Terms per Query:</td>
<td>17.6</td>
<td>6.1</td>
</tr>
</tbody>
</table>
User Study 1: Medical Librarians - Synonyms

- Adding Metathesaurus synonyms doubled Recall without sacrificing Precision.
- WordNet had no influence.
User Study 1: Cancer Researchers - Synonyms

Recall

Precision

• Adding Synonyms did not improve Recall, but it lowered Precision.
User Study 2: Medical Librarians - Related Concepts

Recall

Precision

• Adding Concept Space terms increased Recall.
• Precision did not suffer when Semantic Net was used for filtering.
User Study 2: Cancer Researchers - Related Concepts

Recall

- Adding Concept Space had no effect on Recall or Precision.
There was no difference in performance for Original and Cleaned Natural Language Queries

Medical Librarians:
- provided large Golden Standards
- 14% of the terms could be extracted from the query
- adding synonyms and related concepts doubled recall, without affecting precision

Cancer Researchers:
- provided very small Golden Standards
- 22% of the terms could be extracted from the query
- adding other terms did not increase recall, but lowered precision
System Developments:

HelpfulMED
HelpfulMED on the Web

- Target users: Medical librarians, medical professionals, advanced patients

- One Site, One World

- Medical information is abundant on the Internet

- No Web-based service currently allows users to search all high-quality medical information sources from one site
HelpfulMED Functionalties

- Search among high-quality medical webpages, updated monthly (350K, to be expanded to 1-2M webpages)
- Search all major evidence-based medicine databases simultaneously
- Use Cancer Space (thesaurus) to find more appropriate search terms (1.3M terms)
- Use Cancer Map to browse categories of cancer journal literature (21K topics)
• Spider technology navigates WWW and collects URLs monthly
• UMLS filter and Noun Phraser technologies ensure quality of medical content
• Web pages meeting threshold level of medical phrase content are collected and stored in database
• Index of medical phrases enables efficient search of collection
• Search engine permits Boolean queries and emphasizes exact phrase matching
Evidence-based Medicine Databases

- 5 databases (to be expanded to 12) including:
  - full-text textbook (*Merck Manual of Diagnosis and Therapy*)
  - guidelines and protocols for clinical diagnosis and practice (*National Guidelines Clearinghouse*, NCI’s *PDQ* database)
  - abstracts to journal literature (*CancerLit* database, *American College of Physicians’ journals*)

- Useful for medical professionals and advanced consumers of medical information
• Suggests highly related noun phrases, author names, and NLM Medical Subject Headings

• Phrases automatically transferred to “Search Medical Webpages” for retrieval of relevant documents

• Contains 1.3 M unique terms, 52.6 M relationships

• Document database includes 830,634 CancerLit abstracts
- Multi-layered graphical display of important cancer concepts supports *browsing* of cancer literature

- Document server retrieves relevant documents

- Presents 21,000 topics of documents in 1180 maps organized in 5 layers
Cardiovascular Disease

Coronary heart disease, including heart attacks, is the major cause of death among women in the United States. In general, cigarette smoking doubles the risk. Carbon monoxide slows the transfer of oxygen from the blood to the body. Nicotine increases the heart rate by 15-25 beats per minute, and blood pressure goes up by 15-25 points. When combined with high blood pressure and high blood cholesterol, smoking multiplies the risk of having a heart attack.

Women smokers who also use oral contraceptives are 10 times more at risk for having a heart attack. Additionally, smoking increases the risk for hypertension and brain hemorrhage.

Cancer

There has been a rapid increase in the number of lung cancer deaths among women. Women who are heavy smokers are 24 times more likely to develop lung cancer than those who have never smoked, and lung cancer is now the leading cause of cancer deaths in women, exceeding even breast cancer.

The warning signals of lung cancer are:
- A cough that won’t go away
HelpfulMED search of Evidence-based Databases

CancerNet
A service of the National Cancer Institute

Adrenocortical Carcinoma (PDQ®)
Treatment - Patients

Table of Contents

Description
What is PDQ?
How to use PDQ

General Information
What is cancer of the adrenal cortex?

Stage Information
Stages of cancer of the adrenal cortex
Stage I
Stage II
Stage III
Stage IV
Recurrent

Treatment Option Overview
How cancer of the adrenal cortex is treated
Treatment by stage
Stage I Adrenocortical Carcinoma
Consulting HelpfulMED Cancer Space (Thesaurus)

- Enter search term
- Select relevant search terms
- New terms are posted
- Search again...
- Or find relevant webpages
Browsing HelpfulMED Cancer Map

1. Visual Site Browser
2. Top level map
3. Diagnosis, Differential
4. Brain Neoplasms
5. Brain Tumors
CMedPort: Intelligent Searching for Chinese Medical Information

Yilu Zhou, Jialun Qin, Hsinchun Chen
Outline

- Introduction
- Related Work
- Research Prototype—CMedPort
- Experimental Design
- Experimental Results
- Conclusions and Future Directions
Introduction

• As the second most popular language online, Chinese occupies 12.2% of Internet languages (Global Reach, 2003).
• There are a tremendous amount of medical Web pages provided in Chinese on the Internet.
• Chinese medical information seekers find it difficult to locate desired information, because of the lack of high-performance tools to facilitate medical information seeking.
Internet Searching and Browsing

• The sheer volume of information makes it more and more difficult for users to find desired information (Blair and Maron, 1985).

• When seeking information on the Web, individuals typically perform two kinds of tasks — Internet searching and browsing (Chen et al., 1998; Carmel et al., 1992).
Internet Searching and Browsing

• Internet Searching is “a process in which an information seeker describes a request via a query and the system must locate the information that matches or satisfies the request.” (Chen et al., 1998).

• Internet Browsing is “an exploratory, information seeking strategy that depends upon serendipity” and is “especially appropriate for ill-defined problems and for exploring new task domains.” (Marchionini and Shneiderman, 1988).
Searching Support Techniques

• Domain-Specific Search Engines
  – General-purpose search engines, such as Google and AltaVista, usually result in thousands of hits, many of them not relevant to the user queries.
  – Domain-specific search engines could alleviate this problem because they offer increased accuracy and extra functionality not possible with general search engines (Chau et al., 2002).
Searching Support Techniques

• Meta-Search
  – By relying solely on one search engine, users could miss over 77% of the references they would find most relevant (Selberg and Etzioni, 1995).
  – Meta-search engines can greatly improve search results by sending queries to multiple search engines and collating only the highest-ranking subset of the returns from each one (Chen et al., 2001; Meng et al., 2001; Selberg and Etzioni, 1995).
Browsing Support Techniques

• Summarization—Document Preview
  – Summarization is another post-retrieval analysis technique that provides a preview of a document (Greene et al., 2000).
  – It can reduce the size and complexity of Web documents by offering a concise representation of a document (McDonald and Chen, 2002).
Browsing Support Techniques

• Categorization— Document Overview
  – Document categorization is based on the Cluster Hypothesis: “closely associated documents tend to be relevant to the same requests” (Rijsbergen, 1979).
  – In a browsing scenario, it is highly desirable for an IR system to provide an overview of the retrieved document.
Browsing Support Techniques

• Categorization—Document Overview
  – In Chinese information retrieval, efficient categorization of Chinese documents relies on the extraction of meaningful keywords from text.
  – The mutual information algorithm has been shown to be an effective way to extract keywords from Chinese documents (Ong and Chen, 1999).
Regional Difference among Chinese Users

• Chinese is spoken by people in mainland China, Hong Kong and Taiwan.

• Although the populations of all three regions speak Chinese, they use different Chinese characters and different encoding standards in computer systems.
  – Mainland China: simplified Chinese (GB2312)
  – Hong Kong and Taiwan: traditional Chinese (Big5)
Regional Difference among Chinese Users

- When searching in a system encoded one way, users are not able to get information encoded in the other.
- Chinese medical information providers in all three regions usually keep only information from their own regions.
- Users who want to find information from other regions have to use different systems.
Current Chinese Search Engines and Medical Portals

• Major Chinese Search Engines
  – www.sina.com (China)
  – hk.yahoo.com (Hong Kong)
  – www.yam.com.tw (Taiwan)
  – www.openfind.com.tw (Taiwan)
Current Chinese Search Engines and Medical Portals

• Features of Chinese search engines
  – They have basic Boolean search function.
  – They support directory-based browsing.
  – Some of them (Yahoo and Yam) provide encoding conversion to support cross-regional search.
  – Their content is NOT focused on Medical domain.
  – They only have one version for their own region.
  – They do not have comprehensive functionality to address users need.
Current Chinese Search Engines and Medical Portals

- Chinese medical portals
  - www.999.com.cn (Mainland China)
  - www.medcyber.com (Mainland China)
  - www.trustmed.com.tw (Taiwan)
Current Chinese Search Engines and Medical Portals

• Features of Chinese medical portals
  – Most of them do not have search function.
  – For those who support search function, they maintain a small collection size.
  – Their content is focused on medical domain and covers information about general health, drug, industry, research papers, research conferences, and etc.
  – They only have one version for their own region.
  – They do not have comprehensive functionality to address users need.
Research Prototype — *CMedPort*
The *CMedPort* (http://ai30.bpa.arizona.edu:8080/gbmed) was built to provide medical and health information services to both researchers and the public.

- The main components are: (1) Content Creation; (2) Meta-search Engines; (3) Encoding Converter; (4) Chinese Summarizer; (5) Categorizer; and (6) User Interface.
CMedPort System Architecture

**Front End**
- User Interface
- Chinese Summarizer
- Text Categorizer
- Control Component

**Middleware**
- User query and request
- Result page list
- Summary result
- Folder display
- Request & result page
- Post Analysis
- Request & result pages
- Query
- Converted result pages
- Chinese Encoding Converter (GB2312 ↔ Big5)

**Back End**
- Simplified Chinese Collection (Mainland China)
  - MS SQL Server
- Traditional Chinese Collections (HK & TW)
  - MS SQL Server
- Meta-search Module
- SpidersRUs Toolkit
- Indexing and loading
- Spidering
- The Internet
- Online Search Engines
Input keywords
Select websites from mainland China, Hong Kong and Taiwan
Select search engines from mainland China, Hong Kong and Taiwan
Results are of both simplified Chinese and traditional Chinese
Original encoding of the result Traditional Chinese results have been converted into simplified Chinese
Show simplified Chinese results directly

Chinese Integrated Categorization

Chinese Integrated Analysis

Results from three different regions are categorized
Research Prototype— *CMedPort*

• Content Creation
  – ‘SpidersRUs’ Digital Library Toolkit (http://ai.bpa.arizona.edu/spidersrus/) developed in the AI Lab was used to collect and index Chinese medical-related Web pages.
  – ‘SpidersRUs’
    • The toolkit used a character-based indexing approach. Positional information on the character was captured for phrase search in retrieval phase.
    • It was able to deal with different encodings of Chinese (GB2312, Big5, and UTF8).
    • It also indexed different document formats, including HTML, SHTML, text, PDF, and MS Word.
Research Prototype— CMedPort

• Content Creation
  – The 210 starting URLs were manually selected based on suggestions from medical domain experts.
  – More than 300,000 Web pages were collected and indexed and stored in a MS SQL Server database.
  – They covered a large variety of medical-related topics, from public clinics to professional journals, and from drug information to hospital information.
Research Prototype—*CMedPort*

- **Meta-search Engines**
  - CMedPort “meta-searches” six key Chinese search engines.
    - [www.baidu.com](http://www.baidu.com) -- the biggest Internet search service provider in mainland China;
    - [www.sina.com.cn](http://www.sina.com.cn) -- the biggest general Web portal in mainland China;
    - [hk.yahoo.com](http://hk.yahoo.com) -- the most popular directory-based search engine in Hong Kong;
    - [search2.info.gov.hk](http://search2.info.gov.hk) -- a high quality search engine provided by the Hong Kong government;
    - [www.yam.com](http://www.yam.com) -- the biggest Chinese search engine in Taiwan;
    - [www.sina.com.tw](http://www.sina.com.tw) -- one of the biggest Web portals in Taiwan.
Research Prototype— CMedPort

• Encoding Converter
  – The encoding converter program used a dictionary with 6,737 entries that map between simplified and traditional Chinese characters.
  – The encoding converter enables cross-regional search and addressed the problem of different Chinese character forms.
Research Prototype—CMedPort

- Chinese Summarizer
  - The Chinese Summarizer is a modified version of TXTRACTOR, a summarizer for English documents developed by the AI Lab (McDonald and Chen, 2002).
  - It is based on a sentence extraction approach using linguistic heuristics such as cue phrases, sentence position and statistical analysis.
Research Prototype—*CMedPort*

- **Categorizer**
  - *CMedPort* Categorizer processes all returned results, and key phrases are extracted from their titles and summaries.
  - Key phrases with high occurrences are extracted as folder topics.
  - Web pages that contain the folder topic are included in that folder.
Experimental Design—Objectives

- The user study was designed to
  - compare CMedPort with regional Chinese search engines to study its effectiveness and efficiency in searching and browsing.
  - evaluate user satisfaction obtained from CMedPort in comparison with existing regional Chinese search engines.
Experimental Design—Tasks and Measures

- Two types of tasks were designed: search tasks and browse tasks.
- Search tasks in our user study were short questions which required specific answers.
- We used *accuracy* as the primary measure of effectiveness in searching tasks as follow:

\[
\text{Accuracy} = \frac{\text{number of correct answers given by the subject}}{\text{total number of questions asked}}
\]
Experimental Design—Tasks and Measures

• Each browse task consisted of a topic that defined an information need accompanied by a short description regarding the task and the related questions.

• Theme identification was used to evaluate performance of browse tasks.

\[
\text{Theme precision} = \frac{\text{number of correct themes identified by the subject}}{\text{number of all themes identified by the subject}}
\]

\[
\text{Theme recall} = \frac{\text{number of correct themes identified by the subject}}{\text{number of correct themes identified by expert judges}}
\]
Experimental Design—Tasks and Measures

- Efficiency in both tasks was directly measured by the time subjects spent on the tasks using different systems.
- System usability questionnaires from Lewis (1995) were used to study user satisfaction toward CMedPort and benchmark systems. Subjects rated the systems with a 1-7 score from different perspectives including effectiveness, efficiency, easiness, interface, error recovery ability, and etc.
Experimental Design—Benchmarks

• Existing Chinese medical portals are not suitable for benchmarks; because they do not have good search functionality and they usually only search for their own content.

• Thus, CMedPort was compared with three major commercial Chinese search engines from the three regions:
  – Sina (mainland China)
  – Yahoo HK (Hong Kong)
  – Openfind (Taiwan)
Experimental Design—Subjects

- Forty-five subjects, fifteen from each region, were recruited from the University of Arizona for the experiment.
- Each subject was required to perform 4 search tasks and 8 browse tasks using CMedPort and another benchmark search engine according to his/her origin.
Experimental Design—Experts

• Three graduate students from the Medical School at the University of Arizona, one from each region, were recruited as the domain experts.

• They provided answers for all search and browse tasks and evaluated the answers of subjects.
Experimental Results and Discussions
Experimental Results—Search Tasks

- **Effectiveness**: Accuracy of search tasks
  - CMedPort achieved significantly higher accuracy than Sina.
  - CMedPort achieved comparable accuracy with Yahoo HK and Openfind.

<table>
<thead>
<tr>
<th>Region</th>
<th>System</th>
<th>Accuracy</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainland China</td>
<td>CMedPort</td>
<td>0.91667</td>
<td>0.008046*</td>
</tr>
<tr>
<td></td>
<td>Sina</td>
<td>0.625</td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td>CMedPort</td>
<td>0.9615</td>
<td>0.163094</td>
</tr>
<tr>
<td></td>
<td>Openfind</td>
<td>0.8461</td>
<td></td>
</tr>
<tr>
<td>Hong Kong</td>
<td>CMedPort</td>
<td>0.9285</td>
<td>0.092418</td>
</tr>
<tr>
<td></td>
<td>Yahoo HK</td>
<td>0.8571</td>
<td></td>
</tr>
</tbody>
</table>
Experimental Results—Search Tasks

- Efficiency of search tasks
  - Users spent significantly less time in search tasks using CMedPort than using Sina and Yahoo HK.
  - Users spent comparable time in search tasks using CMedPort and Openfind.

<table>
<thead>
<tr>
<th>Region</th>
<th>System</th>
<th>Time (seconds)</th>
<th>$p$-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainland China</td>
<td>CMedPort</td>
<td>97.962</td>
<td><strong>0.03779</strong></td>
</tr>
<tr>
<td></td>
<td>Sina</td>
<td>149.039</td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td>CMedPort</td>
<td>72.4333</td>
<td>0.0193905</td>
</tr>
<tr>
<td></td>
<td>Openfind</td>
<td>114.7667</td>
<td></td>
</tr>
<tr>
<td>Hong Kong</td>
<td>CMedPort</td>
<td>95.0333</td>
<td><strong>0.044801</strong></td>
</tr>
<tr>
<td></td>
<td>Yahoo HK</td>
<td>117.9667</td>
<td></td>
</tr>
</tbody>
</table>
Experimental Results—Browse Tasks

- Effectiveness: Theme precision of browse tasks
  - CMedPort achieved significantly higher theme precision than Openfind.
  - CMedPort achieved comparable theme precision with Sina and Yahoo HK.

<table>
<thead>
<tr>
<th>Region</th>
<th>System</th>
<th>Theme Precision</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainland China</td>
<td>CMedPort</td>
<td>0.819327</td>
<td>0.071138</td>
</tr>
<tr>
<td></td>
<td>Sina</td>
<td>0.675099</td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td>CMedPort</td>
<td>0.78919</td>
<td>0.031372*</td>
</tr>
<tr>
<td></td>
<td>Openfind</td>
<td>0.636172</td>
<td></td>
</tr>
<tr>
<td>Hong Kong</td>
<td>CMedPort</td>
<td>0.790508</td>
<td>0.05063</td>
</tr>
<tr>
<td></td>
<td>Yahoo HK</td>
<td>0.651905</td>
<td></td>
</tr>
</tbody>
</table>
Experimental Results—Browse Tasks

- Effectiveness: Theme recall of browse tasks
  - CMedPort achieved significantly higher theme recall than all three benchmark systems.

<table>
<thead>
<tr>
<th>Region</th>
<th>System</th>
<th>Theme Recall</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainland China</td>
<td>CMedPort</td>
<td>0.47777</td>
<td>0.000541*</td>
</tr>
<tr>
<td></td>
<td>Sina</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td>CMedPort</td>
<td>0.480769</td>
<td>&lt;0.00001*</td>
</tr>
<tr>
<td></td>
<td>Openfind</td>
<td>0.215385</td>
<td></td>
</tr>
<tr>
<td>Hong Kong</td>
<td>CMedPort</td>
<td>0.524</td>
<td>&lt;0.00001*</td>
</tr>
<tr>
<td></td>
<td>Yahoo HK</td>
<td>0.228</td>
<td></td>
</tr>
</tbody>
</table>
Experimental Results—Browse Tasks

• Efficiency of browse tasks:
  – Users spent significantly less time in browse tasks using CMedPort than using Sina and Openfind.
  – User spent comparable time in browse tasks using CMedPort and Yahoo HK.

<table>
<thead>
<tr>
<th>Region</th>
<th>System</th>
<th>Time (seconds)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainland China</td>
<td>CMedPort</td>
<td>312.9615</td>
<td><strong>0.002948</strong>*</td>
</tr>
<tr>
<td></td>
<td>Sina</td>
<td>412.2308</td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td>CMedPort</td>
<td>218.1</td>
<td><strong>0.00062</strong>*</td>
</tr>
<tr>
<td></td>
<td>Openfind</td>
<td>318.2667</td>
<td></td>
</tr>
<tr>
<td>Hong Kong</td>
<td>CMedPort</td>
<td>360.4</td>
<td>0.29007</td>
</tr>
<tr>
<td></td>
<td>Yahoo HK</td>
<td>376.73</td>
<td></td>
</tr>
</tbody>
</table>
Experimental Results—User Satisfaction

• User satisfaction:
  – CMedPort achieved significantly higher user satisfaction than all three benchmark systems.

<table>
<thead>
<tr>
<th>Region</th>
<th>System</th>
<th>Average Score from 19 questions (scale: 1-7)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainland China</td>
<td>CMedPort</td>
<td>6.077</td>
<td>&lt;0.00001*</td>
</tr>
<tr>
<td></td>
<td>Sina</td>
<td>4.589</td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td>CMedPort</td>
<td>5.773</td>
<td>0.000338*</td>
</tr>
<tr>
<td></td>
<td>Openfind</td>
<td>5.068</td>
<td></td>
</tr>
<tr>
<td>Hong Kong</td>
<td>CMedPort</td>
<td>5.861</td>
<td>0.001729*</td>
</tr>
<tr>
<td></td>
<td>Yahoo HK</td>
<td>5.013</td>
<td></td>
</tr>
</tbody>
</table>
Experimental Results—User Satisfaction

- User satisfaction:
  - Evaluation of CMedPort individual components.

<table>
<thead>
<tr>
<th>Component (scale: 1-7)</th>
<th>Summarization</th>
<th>Categorization</th>
<th>Cross-regional Search</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.665</td>
<td>6.174</td>
<td>5.956</td>
</tr>
</tbody>
</table>
Experimental Results—Verbal Comments

• Users verbal comments
  – CMedPort provided a wide coverage and high quality of information:
    • “Showing results from all three regions was more convenient.”
    • “CMedPort gave more specific answers.”
    • “It is easier to find information from CMedPort.”
    • “CMedPort provides more in-depth information.”
  – Subjects liked summarizer and categorizer
    • “Categorizer is really helpful. It allows me to locate the useful information.”
    • “Summarization is useful when the Web page is long.”
Experimental Results—Verbal Comments

– Users liked the interface of CMedPort
  • “The interface is clear and easy to understand.”
– They suggested other functions and pointed out places for improvement.
  • “I hope to see the key words highlighted in the result description.”
  • “I hope it could be faster.”
  • “The category names are very related to what I’m looking for.”
Discussions

• CMedPort achieved comparable effectiveness with regional Chinese search engines in searching.

• CMedPort achieved comparable theme precision and significantly higher theme recall than regional Chinese search engines in browsing.

• The higher theme recall benefited from
  – High quality of local collection
  – Diverse meta-search engines incorporated
  – Cross-regional search capability
Discussions

• CMedPort achieved comparable efficiency with regional Chinese search engines in both searching and browsing.
• Users’ subjective evaluations on overall satisfaction of CMedPort were higher than those of regional Chinese search engines.
• Users liked the analysis capabilities integrated in CMedPort and the cross-regional search function.
Web Mining: Machine Learning for Web Applications

Hsinchun Chen and Michael Chau
Outline

• Introduction
• Machine Learning: An Overview
• Machine Learning for Information Retrieval: Pre-Web
• Web Mining
• Conclusions and Future Directions
The Web’s **large** size and its **unstructured** and **dynamic** content, as well as its **multilingual** nature make extracting useful knowledge from it a challenging research problem.

- **Machine Learning** techniques can be a possible approach to solve these problems and also **Data Mining** has become a significant subfield in this area.

- The various activities and efforts in this area are referred to as **Web Mining**.
What is Web Mining?

- The term Web Mining was coined by Etzioni (1996) to denote the use of Data Mining techniques to automatically discover Web documents and services, extract information from Web resources, and uncover general patterns on the Web.

- In this article, we have adopted a broad definition that considers Web mining to be “the discovery and analysis of useful information from the World Wide Web” (Cooley et al., 1997).

- Also, web mining research overlaps substantially with other areas, including data mining, text mining, information retrieval, and web retrieval. (See Table 1)
<table>
<thead>
<tr>
<th>Purpose</th>
<th>Any data</th>
<th>Textual data</th>
<th>Web-related data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieving known data or documents efficiently and effectively</td>
<td>Data Retrieval/Database</td>
<td>Information Retrieval</td>
<td>Web Retrieval</td>
</tr>
<tr>
<td>Finding new patterns or knowledge previously unknown to the system</td>
<td>Data Mining</td>
<td>Text Mining</td>
<td>Web Mining</td>
</tr>
</tbody>
</table>

Table 1. A classification of retrieval and mining techniques and applications.
In General, **Machine learning algorithms** can be classified as

- **Supervised learning**: Training examples contain input/output pair patterns. Learn how to predict the output values of new examples.

- **Unsupervised learning**: Training examples contain only the input patterns and no explicit target output. The learning algorithm needs to generalize from the input patterns to discover the output values.

We have identified the following five major Machine Learning paradigms:

- Probabilistic models
- Symbolic learning and rule induction
- Neural networks
- Analytic learning and fuzzy logic.
- Evolution-based models

**Hybrid approaches**: The boundaries between the different paradigms are usually unclear and many systems have been built to **combine** different approaches.
Learning techniques had been applied in Information Retrieval (IR) applications long before the recent advances of the Web.

In this section, we will briefly survey some of the research in this area, covering the use of Machine Learning in

- Information extraction
- Relevance feedback
- Information filtering
- Text classification and text clustering
Web Mining research can be classified into three categories:

- **Web content mining** refers to the discovery of useful information from Web contents, including text, images, audio, video, etc.

- **Web structure mining** studies the model underlying the link structures of the Web.
  
  - It has been used for search engine result ranking and other Web applications (e.g., Brin & Page, 1998; Kleinberg, 1998).

- **Web usage mining** focuses on using data mining techniques to analyze search logs to find interesting patterns.
  
  - One of the main applications of Web usage mining is its use to learn user profiles (e.g., Armstrong et al., 1995; Wasfi et al., 1999).
Text Mining for Web Documents

Text mining for Web documents can be considered a sub-field of **Web content mining**.

- **Information extraction techniques** have been applied to Web HTML documents
  - E.g., *Chang and Lui (2001)* used a PAT tree to construct automatically a set of rules for information extraction.

- **Text clustering algorithms** also have been applied to Web applications.
  - E.g., *Chen et al. (2001; 2002)* used a combination of noun phrasing and SOM to cluster the search results of search agents that collect Web pages by meta-searching popular search engines.
Web Spiders, have been defined as “software programs that traverse the World Wide Web by following hypertext links and retrieving Web documents by HTTP protocol” (Cheong, 1996).

They can be used to
- build the databases of search engines (e.g., Pinkerton, 1994)
- perform personal search (e.g., Chau et al., 2001)
- archive Web sites or even the whole Web (e.g., Kahle, 1997)
- collect Web statistics (e.g., Broder et al., 2000)

Intelligent Web Spiders: some spiders that use more advanced algorithms during the search process have been developed.
- E.g., the Itsy Bitsy Spider searches the Web using a best-first search and a genetic algorithm approach (Chen et al., 1998a).
In order to extract non-English knowledge from the web, Web Mining systems have to deal with issues in language-specific text processing.

The base algorithms behind most machine learning systems are language-independent. Most algorithms, e.g., text classification and clustering, need only to take a set of features (a vector of keywords) for the learning process.

However, the algorithms usually depend on some phrase segmentation and extraction programs to generate a set of features or keywords to represent Web documents.

Other learning algorithms such as information extraction and entity extraction also have to be tailored for different languages.
Web Visualization

- **Web Visualization** tools have been used to help users maintain a "big picture" of the retrieval results from search engines, web sites, a subset of the Web, or even the whole Web.
  - The most well known example of using the tree-metaphor for Web browsing is the hyperbolic tree developed by Xerox PARC (*Lamping & Rao, 1996*).

- In these visualization systems, Machine Learning techniques are often used to determine how Web pages should be placed in the 2-D or 3-D space.
  - One example is the SOM algorithm described earlier (*Chen et al., 1996*).
The Semantic Web

- **Semantic Web** technology (*Berners-Lee et al., 2001*) tries to add **metadata** to describe data and information on the Web. Based on standards like RDF and XML.

- **Machine learning** can play three roles in the Semantic Web:
  - First, machine learning can be used to **automatically** create the markup or metadata for existing unstructured textual documents on the Web.
  - Second, machine learning techniques can be used to create, merge, update, and maintain **Ontologies**.
  - Third, machine learning can understand and perform **reasoning** on the metadata provided by the Semantic Web in order to extract knowledge from the Web more effectively.
Web Structure Mining

- **Web link structure** has been widely used to infer important web pages information.

- Web structure mining has been largely influenced by research in
  - **Social network analysis**
  - **Citation analysis** (bibliometrics).
    - *in-links*: the hyperlinks pointing to a page
    - *out-links*: the hyperlinks found in a page.
    - Usually, the larger the number of in-links, the **better** a page is.

- By analyzing the pages containing a **URL**, we can also obtain
  - **Anchor text**: how other Web page authors annotate a page and can be useful in predicting the content of the target page.
• Web structure mining algorithms:

  – The PageRank algorithm is computed by weighting each in-link to a page proportionally to the quality of the page containing the in-link (Brin & Page, 1998).

  – The qualities of these referring pages also are determined by PageRank. Thus, a page $p$ is calculated recursively as follows:

  $$\text{PageRank}(p) = (1 - d) + d \times \sum_{\text{all } q \text{ linking to } p} \left( \frac{\text{PageRank}(q)}{c(q)} \right)$$

  where $d$ is a damping factor between 0 and 1, $c(q)$ is the number of out-going links in a page $q$. 
Web Structure Mining Algorithms

- Web structure mining algorithms:
  - *Kleinberg (1998)* proposed the **HITS** (Hyperlink-Induced Topic Search) algorithm, which is similar to PageRank.
    
    - **Authority pages**: high-quality pages related to a particular search query.
    - **Hub pages**: pages provide pointers to other authority pages.
    - A page to which many others point should be a good authority, and a page that points to many others should be a good hub.

\[
\text{AuthorityScore}(p) = \sum_{all \ q \ linking \ to \ p} (\text{HubScore}(q))
\]

\[
\text{HubScore}(p) = \sum_{all \ r \ linking \ from \ p} (\text{AuthorityScore}(r))
\]
Another application of Web structure mining is to understand the structure of the Web as a whole.

The core of the Web is a strongly connected component and that the Web’s graph structure is shaped like a bowtie. Broder et al. (2000)

- Strongly Connected Component (SCC); 28% of the Web.
- IN: every Web page contains a direct path to the SCC; 21% of Web
- OUT: a direct path from SCC linking to it; 21% of Web
- TENDRILS: pages hanging off from IN and OUT but without direct path to SCC; 22% of Web
- Isolated, Disconnected Components that are not connected to the other 4 groups; 8% of Web
Web Usage Mining

- Web servers, Web proxies, and client applications can quite easily capture Web Usage data.
  - Web server log: Every visit to the pages, what and when files have been requested, the IP address of the request, the error code, the number of bytes sent to user, and the type of browser used…

- By analyzing the Web usage data, web mining systems can discover useful knowledge about a system’s usage characteristics and the users’ interests which has various applications:
  - Personalization and Collaboration in Web-based systems
  - Marketing
  - Web site design and evaluation
  - Decision support (e.g., Chen & Cooper, 2001; Marchionini, 2002).
Pattern Discovery and Analysis

• Web usage mining has been used for various purposes:
  – **A knowledge discovery process** for mining marketing intelligence information from Web data. *Buchner and Mulvenna (1998)*
  – **Web traffic patterns** also can be extracted from Web usage logs in order to improve the performance of a Web site (*Cohen et al.*, 1998).
  – **Commercial products**: *Web Trends* developed by NetIQ, *WebAnalyst* by Megaputer and *NetTracker* by Sane Solutions.

• Search engine transaction logs also provide valuable knowledge about **user behavior** on Web searching.

• Such information is very useful for a better understanding of users’ Web searching and information seeking behavior and can improve the design of Web search systems.
One of the major goals of Web usage mining is to reveal **interesting trends and patterns** which can often provide important knowledge about the users of a system.

The **Framework** for Web usage mining. *Srivastava et al. (2000)*

- **Preprocessing:** Data cleansing
- **Pattern discovery:** Generic machine learning and Data mining
- **Pattern analysis:** techniques, such as association rule mining, classification, and clustering, often can be applied.

For instance, *Yan et al. (1996)* performed clustering on Web log data to identify users who have accessed similar Web pages.
Many Web applications aim to provide personalized information and services to users. Web usage data provide an excellent way to learn about users’ interest (Srivastava et al., 2000).

- WebWatcher (Armstrong et al., 1995)
- Letizia (Lieberman, 1995)

Web usage mining on Web logs can help identify users who have accessed similar Web pages. The patterns that emerge can be very useful in collaborative Web searching and filtering.

- Amazon.com uses collaborative filtering to recommend books to potential customers based on the preferences of other customers having similar interests or purchasing histories.
- Huang et al. (2002) used Hopfield Net to model user interests and product profiles in an online bookstore in Taiwan.
Conclusions and Future Directions

• Most Web mining activities are still in their early stages and should continue to develop as the Web evolves.

• Future research directions:
  – Multimedia data mining: a picture is worth a thousand words.
  – Multilingual knowledge extraction: Web page translations
  – Wireless Web: WML and HDML.
  – The Hidden Web: forms, dynamically generated Web pages.
  – Semantic Web

• We believe that research in Machine learning and Web mining will help develop applications that can more effectively and efficiently utilize the Web of knowledge of the humankind.
Web Programming with Amazon, Google, and eBay APIs

Chunju Tseng (Lu)
Outline

- Web Services in a nutshell
- AJAX
- Overview of Amazon, Google, and Ebay APIs
- More Web Services and Mashup
- Lab Session: Web Services Programming
What is Web Services?

• Web Services:
  – A new way of reuse/integrate third party software or legacy system
  – No matter where the software is, what platform it residents, or which language it was written in
  – Based on XML and Internet protocols (HTTP, SMTP…)

• Benefits:
  – Ease of integration
  – Develop applications faster
Web Services Architecture

- Simple Object Access Protocol (SOAP)
- Web Service Description Language (WSDL)
- Universal Description, Discovery and Integration (UDDI)
New Breeds of Web Services

• Representational State Transfer (REST)
  – Use HTTP Get method to invoke remote services (not XML)
  – The response of remote service can be in XML or any textual format
  – Benefits:
    • Easy to develop
    • Easy to debug (with standard browser)
    • Leverage existing web application infrastructure
Server Responses in REST

- Really Simple Syndication (RSS, Atom)
  - XML-based standard
  - Designed for news-oriented websites to “Push” content to readers
  - Excellent to monitor new content from websites
- JavaScript Object Notation (JSON)
  - Lightweight data-interchange format
  - Human readable and writable and also machine friendly
  - Wide support from most languages (Java, C, C#, PHP, Ruby, Python…)
JSON Example

```json
{"menu": {
    "id": "file",
    "value": "File",
    "popup": {
        "menuitem": [
            {
                "value": "New",
                "onclick": "CreateNewDoc()"
            },
            {
                "value": "Open",
                "onclick": "OpenDoc()"
            },
            {
                "value": "Close",
                "onclick": "CloseDoc()"
            }
        ]
    }
}}
```

The same text expressed as **XML**:

```xml
<menu id="file" value="File">
    <popup>
        <menuitem value="New" onclick="CreateNewDoc()" />
        <menuitem value="Open" onclick="OpenDoc()" />
        <menuitem value="Close" onclick="CloseDoc()" />
    </popup>
</menu>
```

http://www.json.org/example.html
Rich Interactivity Web - AJAX

• **AJAX**: Asynchronous JavaScript + XML
  • **AJAX incorporates:**
    – standards-based presentation using *XHTML and CSS*;
    – dynamic display and interaction using the *Document Object Model*;
    – data interchange and manipulation using *XML and XSLT*;
    – asynchronous data retrieval using *XMLHttpRequest*;
    – and *JavaScript* binding everything together.

• **Examples:**
  – [http://www.gmail.com](http://www.gmail.com)
  – [http://www.kiko.com](http://www.kiko.com)

AJAX Application Model

Classic web application model:
- User Interface
- HTTP request
- HTML+CSS data
- Web server
- Datastores, backend processing, legacy systems

Ajax web application model:
- User Interface
- JavaScript call
- HTML+CSS data
- Ajax engine
- HTTP request
- XML data
- Web and/or XML server
- Datastores, backend processing, legacy systems
Amazon Web Services (AWS)

- Amazon E-Commerce Service
  - Search catalog, retrieve product information, images and customer reviews
  - Retrieve wish list, wedding registry…
  - Search seller and offer

- Alexa Services
  - Retrieve information such as site rank, traffic rank, thumbnail, related sites amount others given a target URL

- Amazon Historical Pricing
  - Programmatic access to over three years of actual sales data

- Amazon Simple Queue and Storage Service
  - A distributed resource manager to store web services results

- Amazon Elastic Compute Cloud
  - Sell computing capacity by the amount you use
Google Web APIs

• Google has a long list of APIs
  – http://code.google.com/apis/

• Google Search
  – AJAX Search API
  – SOAP Search API (deprecated)
  – Custom search engine with Google Co-op

• Google Map API

• Google Data API (GData)
  – Blogger, Google Base, Calendar, Gmail, Spreadsheets, and a lot more

• Google Talk XMPP for communication and IM

• Google Translation (http://www.oreillynet.com/pub/h/4807)

• Many more undocumented/unlisted APIs to be discovered in Google Blog
eBay API

• Buyers:
  – Get the current list of eBay categories
  – View information about items listed on eBay
  – Display eBay listings on other sites
  – Leave feedback about other users at the conclusion of a commerce transaction

• Sellers:
  – Submit items for listing on eBay
  – Get high bidder information for items you are selling
  – Retrieve lists of items a particular user is currently selling through eBay
  – Retrieve lists of items a particular user has bid on
Other Services/APIs Providers

• Yahoo!  http://developer.yahoo.com/
  – Search (web, news, video, audio, image…) 
  – Flickr, del.icio.us, MyWeb, Answers API

  – Search (SOAP, REST)
  – Spaces (blog), Virtual Earth, Live ID

• Wikipedia
  – Downloadable database

• Many more at Programmableweb.com
  – http://www.programmableweb.com/apis
Services by Category

- **Search**
  - Google, MSN, Yahoo

- **E-Commerce**
  - Amazon, Ebay, Google Checkout
  - TechBargain, DealSea, FatWallet

- **Mapping**
  - Google, Yahoo!, Microsoft

- **Community**
  - Blogger, MySpace, MyWeb
  - del.icio.us, StumbleUpon

- **Photo/ Video**
  - YouTube, Google Video, Flckr

- **Identity/ Authentication**
  - Microsoft, Google, Yahoo

- **News**
  - Various news feed websites including Reuters, Yahoo! and many more.
Mashup
A Novel Form of Web Reuse

• “A mashup is a website or application that combines content from more than one source into an integrated experience.” – Wikipedia

• API X + API Y = mashup Z

• Business model: Advertisement
Mashup: Weather Bonk

- [http://www.weatherbonk.com/weather/index.jsp](http://www.weatherbonk.com/weather/index.jsp)

  - APIs: Google AdWords + Google Maps + hostip.info + MS Virtual Earth + NASA + NOAA Weather Service + WeatherBug + Yahoo Geocoding + Yahoo Maps + Yahoo Traffic
Lab Session:
Amazon, Google, EBay APIs

• Required Software:
  – Eclipse IDE V 3.21 (http://eclipse.org)
  – Amazon API (http://www.amazon.com/gp/aws/landing.html)
  – Amazon API Playground (http://awszone.com)
Setup NetBeans with Amazon

- Install Eclipse IDE for Java
- Register Amazon subscription ID
- Download Amazon Java sample
- Create project
- Install necessary libraries
- Execute Amazon code sample
  - Main.java
Retrieve Customer Information

- **CustomerContentSearch**
  - Give name to retrieve customer ID

- **CustomerContentLookup**
  - Give customer ID
  - Request WishList information
  - ResponseGroup=CustomerInfo, CustomerLists
  - Receive WishList ID

- Use AWSzone.com!
Retrieve Customer WishList and Response Manipulation

- **ListLookup**
  - Give list ID and type to retrieve WishList content with a simple browser (REST)
  - ResponseGroup=ListItems,ListInfo
  - Result shows in XML format
- **ListLookup with xml-to-text XSLT**
  - Give custom XSL to manipulate response
  - Style=http://www.u.arizona.edu/~chunju/text.xsl&Content Type=text/plain
- **ListLookup with xml-to-html XSLT**
  - ResponseGroup=ListItems,ListInfo,Images,SalesRank,ItemAttributes
  - Style=http://www.u.arizona.edu/~chunju/html.xsl&Content Type=text/html
Access Amazon REST Service with Jakarta HttpClient

- Download & install necessary libraries
  - Jakarta HttpClient, Codec
  - JUnit
- Write first Java program in NetBeans
- Use previous REST invocation strings
- Print Amazon response in console
- Resources:
  - http://www.junit.org/
  - http://www.u.arizona.edu/~chunju/HttpClientMain.java
Other Resources

- Java Web Services Developer Pack

- Consuming Web Services with NetBeans

- XSLT Tutorial
  - http://www.w3schools.com/xsl/
Using Open Web APIs in Teaching Web Mining

Hsinchun Chen¹
Xin Li¹
Michael Chau²
Yi-Jen Ho¹
Chunju Tseng¹
Outline

• Introduction
• Background
• A Class Project Combining Web Mining and Open Web APIs
• Student Projects
• Conclusions and Future Directions
• Acknowledgement
Introduction

• The World Wide Web has become an indispensable part of many business organizations.

• In order to effectively utilize the power of the Web, information technology (IT) professionals need to have sufficient knowledge and experience in various Web technologies and applications.

• Traditional education in computer science and information systems curricula may address these areas to some extent, but is definitely not sufficient.

• In recent years, new courses in Internet- and Web-related topics have been offered in many universities around the world to better equip students with such knowledge.
Introduction

• Besides the basic courses, such as Internet networking, Internet application development, and Web search engines, more advanced topics, such as Web mining, are also becoming increasingly important.

• Web mining, defined as the use of data mining, text mining, and information retrieval techniques to extract useful patterns and knowledge from the Web (Etzioni, 1996; Chen & Chau, 2004), has been frequently used in real world applications, such as
  – business intelligence (Chau et al., 2002),
  – Website design (Fang et al., 2006),
  – and customer opinion analysis (Liu et al., 2005).

• However, building a Web mining application or a Web services application from scratch is not an easy task that every student could complete in a semester.
Web Mining

- Since the advent of the Internet, many studies have investigated the possibility of extracting knowledge and patterns from the Web because it is publicly available and contains a rich set of resources.

- Many Web mining techniques are adopted from data mining, text mining, and information retrieval research (Etzioni, 1996; Chen & Chau, 2004).

- Most of these studies aimed to discover resources, patterns, and knowledge from the Web and Web-related data.
Web Mining

- Web mining research can be classified into three categories: Web content mining, Web structure mining, and Web usage mining (Kosala & Blockeel, 2000).
  - Web content mining refers to the discovery of useful information from Web contents, including text, images, audio, video, etc.
  - Web structure mining studies the model underlying the link structures of the Web.
  - Web usage mining focuses on using data mining techniques to analyze search logs or other activity logs to find interesting patterns.
Open Web APIs

• The Web no longer contains merely static pages and contents. Powered by modern database technologies, network technologies, and computational ability developments, many Websites provide sophisticated services to customers.

• By viewing the service provided by a Website as an application (Baresi et al., 2000), the different Website functionalities can be considered different modules of the application.

• The concept of Web APIs enables direct access to these modules from the client side or a third party’s Website.
  
  - Currently, more than 200 Websites have published Web APIs for access to their services (ProgrammableWeb, 2006).
  
  - The idea is to leverage third party efforts on value-adding services and GUIs, so that the Website can focus on and enlarge the usage of its core service.
Open Web APIs

• Many of the open Web APIs are constructed based on the Web services architecture. Some of them have been wrapped into libraries written in Java, .NET, JavaScript, etc., which hide the detailed Web services protocols from the developers and make it easier for the developers to use.

• Web services are based on XML and HTTP, which provide better platform and language independent interoperability.

• Implementation details of the server side are hidden from the users/third party developers, but the interfaces are publicly available. The structured data acquired from the Web application can also be used for further data mining analysis.
Open Web APIs

• The companies that open Web APIs to the public belong to several different categories, such as
  – Web search (e.g., Google and Yahoo),
  – chat and messaging (e.g., MSN and AOL),
  – geographic map (e.g., NASA, Google Maps, Yahoo Map, and Microsoft MapPoint),
  – e-commerce (e.g., Amazon, eBay, and Paypal),
  – shipping (e.g., FedEx and UPS),
  – and others (e.g., BBC and Skype).

• Three sets of open Web APIs that are most relevant to the current paper
  – Amazon
  – eBay
  – Google
As a leader in e-commerce, Amazon Web Services offers a variety of Web services that allow developers to build businesses based on Amazon’s data. Such Web APIs include

- access to Amazon’s product data and e-commerce functionality (Amazon E-Commerce Service),
- access to Amazon’s sales history data (Amazon Historical Pricing),
- access of storage (Amazon S3), queue (Amazon Simple Queue Service), and Web search (Alexa Web Search Platform) (Amazon, 2006).

Amazon E-Commerce Service is based on SOAP (Simple Object Access Protocol) and WSDL (Web Service Description Language) standards (Curbera et al., 2002).

The Amazon Historical Pricing service gives developers programmatic access to the actual sales data for books, music, videos, and DVDs (as sold by third party sellers on Amazon.com) since 2002.
eBay

• Using the eBay Web Services API, developers can create Web-based applications to conduct business with the eBay Platform (Mueller, 2004).
  – The API can access the data on eBay.com and Half.com.
  – Developers can perform functions such as sales management, item search, and user account management (eBay, 2006).

• The eBay Web Services API is also based on the SOAP and WSDL standards.

• The eBay developer center provides wrappers of the SOAP APIs in Java, .NET, and PHP. To use the eBay API, developers need to join the eBay developer program (free).
Google

- Google provides several kinds of open Web APIs for accessing their
  - advertisement service (AdWords API),
  - blog service (Blogger Atom API),
  - map service (Google Maps API),
  - and the Web search service (Google Search API).

- The most widely used are Google Search API and Google Maps API (Google, 2006).
  - Google Search API are implemented as a Web service using SOAP and WSDL standards. Google provides a Java library which wraps the SOAP APIs.
  - Google Maps API is a JavaScript API which can embed Google Maps in Web pages. It is designed mainly for data representation purposes.
A Class Project Combining Web Mining and Open Web APIs

• Given the advantages of using open Web APIs and the vast amount of Web APIs provided by different Websites, we found few course projects that consolidated open Web APIs within Web mining projects.

• We believe that it is important for students to acquire some classroom knowledge and experiences in both areas.

• In this paper, we report on a class project that we designed and developed in 2005.
A Class Project Combining Web Mining and Open Web APIs

• In the class project each group of students was required to create a Web business with a complete Website and business functionalities for specific customers, using one or more of the three open Web APIs (Amazon, eBay, and Google) together with other data mining techniques learned in the class.

• Given the comprehensive functions provided by the three Web APIs, the main challenge of the project was to integrate these components and design attractive system features.

• Students were encouraged to integrate open source data mining components, such as WEKA (Witten & Frank, 2005), to analyze the data retrieved through the Web APIs.
A Class Project Combining Web Mining and Open Web APIs

- Each team consisted of three members who would participate in the design, coding, implementation, and analysis of the prototype.
  - Each team member was required to participate in all aspects of the project.
  - They could use their own machines as a server or share a server provided by the university’s open computing laboratory.

- Lab sessions were provided to familiarize students with Web API programming. A teaching assistant was also provided to assist the students with programming problems.

- The students were given three and a half months to finish the project and were required to submit a project proposal after the first month of class.
  - At the end of the semester, each group would present their work, demonstrate the system and submit a final report.
  - The projects were graded based on system functionalities, novelty, and business feasibility.
Student Projects

• At the end of the semester, each project group gave a demonstration of their system.

• Overall, we found the projects well designed and interesting.
  – The students applied what they learned from the class as well as demonstrating their creativity and teamwork in the project.
  – By incorporating the services and data provided by Amazon, Google, and eBay with the latest data mining and visualization technologies, the students developed many innovative business models.

• Four selected examples
  – Wishsky
  – Tucson Book Xchange
  – Cellphone Intelligent Auctioning (CIA)
  – SciBubble
Wishsky

- Wishsky is an integrated “wish list” management system which enables customers to monitor the news and ongoing sales/auctions related to the products in their wish lists. Wishsky also recommends alternative products and related products to the customers.

- Wishsky was implemented on Tomcat using Java Server Page (JSP).

- The system architecture has three layers.
  - The bottom layer of the system is the database layer. Wishsky uses a local Microsoft SQL Server which stores the customer information, wish list data, and product recommendation information through Java Database Connectivity (JDBC).
  
  - The middle layer of the system is the logic control layer, which contains four modules (written in Java). The API Query Module sends requests to and gets results form eBay, Google, and Amazon.
  
  - The top layer of the system is the presentation layer. HTML, JSP, and JavaScript are used to generate friendly and intuitive user interfaces.
Wishsky System Architecture
Wishsky

Popular products' report

Apple 4 GB iPod Nano
Black

Star Wars - Episode I,
The Phantom Menace
(Widescreen Edition)

Star Wars, Episode III - Revenge of the Sith
(Widescreen Edition)

News

Target: Entertainment: Star Wars, Episode III - Revenge of the Sith (Widescreen Edition) [DVD].
$17.98. List price: $29.98. You Save: $12.00 (40%). Quantity: ...

Star Wars: Episode III Revenge of the Sith Movie news, stills...

Star Wars: Episode III Revenge of the Sith database entry on Monsters and...

Star Wars: Battlefront (Gaming). Star Wars Trilogy (Widescreen Edition) (DVD) ...

Star Wars: Episode III, Revenge of the Sith - Movie Trailers...

More News about The DVD From The Official Star Wars Web ...


Star Wars: Episode III - Revenge of the Sith - Cast, Crew, Reviews, Plot

Wishsky Website
A Sample Scenario for Wishsky

- The Wishsky homepage features the most popular items in the Wishsky system according to users’ wish lists. In addition to the item descriptions (e.g., “Star Wars” and “Episode III”), the right frame shows news related to those popular items, which was retrieved using the Google Search API.
A Sample Scenario for Wishsky

- After registration, the user can create a Wishsky wish list by importing and modifying his/her Amazon wish list.

The page for personalizing a wish list
A Sample Scenario for Wishsky

• After the wish list is set up, the system will personalize each user’s view to include the wish list, item-related news, and product recommendations. The user can choose among different algorithms to obtain product recommendations or search available auctions of their wish list items on eBay.
Tucson Book Xchange

• Tucson Book Xchange is an online book exchange system.
  – The system supports online book selling and buying, mainly for Tucson residents.
  – It also enables users to search for books on Amazon and eBay Websites by integrating their Web APIs.

• One unique feature of the system is that it visualizes the actual location of the books.
  – By incorporating Google Maps API, the system is able to provide map images and driving directions according to buyer’s and seller’s home addresses in their profiles.
An Example for Tucson Book Xchange

- After user “aydink” reserved a book, the map image in the right frame locates the user and the book, and the directions and address information are shown in the left frame. The Tucson Book Xchange system architecture is similar to the Wishsky system.

<table>
<thead>
<tr>
<th>User</th>
<th>Request Date</th>
<th>Status</th>
<th>Due Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>aydink</td>
<td>2005-12-01</td>
<td>Using</td>
<td>2005-12-15</td>
</tr>
</tbody>
</table>

Total Time: 7 mins

The user interface of Google Maps API integration in Tucson Book Xchange
Cellphone Intelligent Auctioning

- Cellphone Intelligent Auctioning (CIA) is a cell phone auction history analysis system which provides value-added services to eBay cell phone buyers and sellers.

- The group developed a standalone application in C# to extract eBay item information and transaction information, which were stored in a backend MySQL database.

- Based on these transaction data, CIA performs analysis on how product price and quantity associate with four product aspects, namely time, brand, seller, and location. The results are presented in tables and graphs which provide users with a better idea of the overall trend of the eBay cell phone market.
Cellphone Intelligent Auctioning

CIA System Architecture
Cellphone Intelligent Auctioning

- The time trend analysis of the average cell phone price for a period of two weeks. The analysis clearly shows the weekend effect, when the bidding prices of items are slightly lower than the average prices. Such analysis may help buyers and sellers make better auction decisions.

The time trend analysis on average cell phone price
To apply geographic analysis on the eBay auction transactions, the group combined several APIs together. After extracting location information from transactions, the graphical coordinates (latitude and longitude) information is retrieved from Yahoo through the Yahoo Geocoding API and fed into the Google Maps API to visualize the locations on a map.
SciBubble

• SciBubble is an Amazon science fiction book portal which features distinct visualization that helps customers find their books of interest.

• The science fiction book data, including book details and customer reviews, were retrieved from Amazon through the Amazon E-Commerce Service API and were loaded into a MySQL database.

• The similarity between each pair of books was pre-calculated according to such information as publisher, rating, publication year, ranking, and category.
SciBubble System Architecture

SciBubble

User Interface (CSS, HTML, JavaScript)

Web Server (Apache, PHP, Applet)

Amazon Database

Data Extractor (Java)

Amazon Database

Data mining Agent (Java)

Main Web Application Architecture
An algorithm was designed to visualize similar books which are related to a given book search query. In this algorithm, each book is represented as a bubble and the similarities between books are represented by the distances and angles between the bubbles.

The visualization interface of SciBubble
# A Summary of Selected Projects and Web APIs

<table>
<thead>
<tr>
<th>Project name</th>
<th>Business model</th>
<th>Web APIs</th>
<th>Innovations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellphone Intelligent Auctioning</td>
<td>Cell phone sales analysis</td>
<td>eBay API, Yahoo Geocoding API, Google Maps API</td>
<td>Statistical analysis</td>
</tr>
<tr>
<td>GiftChannel</td>
<td>Wish list management</td>
<td>Amazon E-Commerce Service</td>
<td>Product recommendation</td>
</tr>
<tr>
<td>MusicBox</td>
<td>Music album sales/news portal</td>
<td>Google Search API, Amazon E-Commerce Service</td>
<td>Product recommendation and visualization</td>
</tr>
<tr>
<td>PriceSmart</td>
<td>Market analysis /Sales management</td>
<td>Amazon E-Commerce Service</td>
<td>Data analysis (SQL server 2005)</td>
</tr>
<tr>
<td>SciBubble</td>
<td>Science fiction book portal</td>
<td>Amazon E-Commerce Service</td>
<td>Visualization</td>
</tr>
<tr>
<td>Tucson Book Xchange</td>
<td>Local book flea market</td>
<td>Amazon E-Commerce Service, eBay API, Google Maps API</td>
<td>Driving directions</td>
</tr>
<tr>
<td>Wishsky</td>
<td>Wish list management</td>
<td>Amazon E-Commerce Service, eBay API, Google Search API</td>
<td>Product recommendation (WEKA)</td>
</tr>
</tbody>
</table>
# Online Resources Used in the Projects

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web APIs</strong></td>
<td>Amazon Web Services</td>
<td><a href="http://www.amazon.com/webservices/">http://www.amazon.com/webservices/</a></td>
</tr>
<tr>
<td></td>
<td>Google APIs</td>
<td><a href="http://code.google.com/apis.html">http://code.google.com/apis.html</a></td>
</tr>
<tr>
<td></td>
<td>eBay API</td>
<td><a href="http://developer.ebay.com/">http://developer.ebay.com/</a></td>
</tr>
<tr>
<td></td>
<td>Yahoo APIs</td>
<td><a href="http://developer.yahoo.com/">http://developer.yahoo.com/</a></td>
</tr>
<tr>
<td><strong>Data analysis packages</strong></td>
<td>WEKA</td>
<td><a href="http://www.cs.waikato.ac.nz/ml/weka/">http://www.cs.waikato.ac.nz/ml/weka/</a></td>
</tr>
<tr>
<td></td>
<td>Yale</td>
<td><a href="http://sourceforge.net/projects/yale/">http://sourceforge.net/projects/yale/</a></td>
</tr>
<tr>
<td></td>
<td>MS SQL Server 2005</td>
<td><a href="http://www.microsoft.com/sql/">http://www.microsoft.com/sql/</a></td>
</tr>
<tr>
<td></td>
<td>IBM DB2 intelligent miner</td>
<td><a href="http://www.ibm.com/software/data/iminer/">http://www.ibm.com/software/data/iminer/</a></td>
</tr>
<tr>
<td><strong>Network visualization toolkits</strong></td>
<td>JUNG</td>
<td><a href="http://jung.sourceforge.net/">http://jung.sourceforge.net/</a></td>
</tr>
<tr>
<td></td>
<td>Graphviz</td>
<td><a href="http://www.graphviz.org/">http://www.graphviz.org/</a></td>
</tr>
</tbody>
</table>
Conclusions and Future Directions

• In this paper, we report on our experience designing a class project for students in a graduate course to use open Web APIs for developing Web mining applications.

• Overall, the results are encouraging.
  – Through these Web mining projects, we observe that most students were able to build innovative Web applications within a short period of time; which would be impossible if the students had to develop the systems from scratch.
  – We observe that students developed the abilities to create interesting business models and integrate necessary system components to implement them.

• In addition to the three suggested APIs from Google, eBay, and Amazon, students also successfully identified and incorporated other Web APIs and data mining/visualization tools in the projects.
Computer-Assisted “Dark Web” Collection, Analysis and Visualization

The Dark Web team
Dark Web

- 2002-, ITIC, NSF, LOC
- Discussions: FBI, DOD/Dept of Army, NSA, DHS
- Connection:
  - Web site spidering
  - Forum spidering
  - Video spidering
- Analysis and Visualization:
  - Link and content analysis (web sites)
  - Web metrics analysis (web sites sophistication)
  - Authorship analysis (forums; CyberGate)
  - Sentiment analysis (forums; CyberGate)
  - Video coding and analysis (videos; MCT)
## Organizations that Capture and Analyze Middle Eastern Terrorists’ Web sites

<table>
<thead>
<tr>
<th>Organization</th>
<th>Description</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Archive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Internet Archive (IA), 1996-</td>
<td>Spidering (every 2 mths.) to collect open access HTML pages</td>
<td></td>
</tr>
<tr>
<td><strong>Research Center</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Artificial Intelligence (AI) Lab, UA, 2003-</td>
<td>Spidering (every 3 mths.) to collect U.S. Domestic, Latin America, and Middle Eastern terrorist groups’ Web sites, forums &amp; multimedia</td>
<td></td>
</tr>
<tr>
<td>3. Anti-terrorism Coalition (ATC), 2003-</td>
<td>Jihad Watch. Has 448 terrorist Web sites &amp; forums</td>
<td></td>
</tr>
<tr>
<td>5. MEMRI, 2003-</td>
<td>Jihad &amp; Terrorism Studies Project.</td>
<td></td>
</tr>
<tr>
<td><strong>Vigilante Community</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dynamic files (e.g., PHP, ASP, JSP, etc.) are widely used in terrorist Web sites, indicating a high level of technical sophistication. Multimedia is also heavily used in terrorist Web sites.

<table>
<thead>
<tr>
<th>Terrorist Collection</th>
<th># of Files</th>
<th>Volume(Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>222,687</td>
<td>12,362,050,865</td>
</tr>
<tr>
<td>Indexable Files</td>
<td>179,223</td>
<td>4,854,971,043</td>
</tr>
<tr>
<td>HTML Files</td>
<td>44,334</td>
<td>1,137,725,685</td>
</tr>
<tr>
<td>Word Files</td>
<td>278</td>
<td>16,371,586</td>
</tr>
<tr>
<td>PDF Files</td>
<td>3,145</td>
<td>542,061,545</td>
</tr>
<tr>
<td>Dynamic Files</td>
<td>130,972</td>
<td>3,106,537,495</td>
</tr>
<tr>
<td>Text Files</td>
<td>390</td>
<td>45,982,886</td>
</tr>
<tr>
<td>Powerpoint Files</td>
<td>6</td>
<td>6,087,168</td>
</tr>
<tr>
<td>XML Files</td>
<td>98</td>
<td>204,678</td>
</tr>
<tr>
<td>Multimedia Files</td>
<td>35,164</td>
<td>5,915,443,276</td>
</tr>
<tr>
<td>Image Files</td>
<td>31,691</td>
<td>525,986,847</td>
</tr>
<tr>
<td>Audio Files</td>
<td>2,554</td>
<td>3,750,390,404</td>
</tr>
<tr>
<td>Video Files</td>
<td>919</td>
<td>1,230,046,468</td>
</tr>
<tr>
<td>Archive Files</td>
<td>1,281</td>
<td>483,138,149</td>
</tr>
<tr>
<td>Non-Standard Files</td>
<td>7,019</td>
<td>1,108,499,397</td>
</tr>
</tbody>
</table>

Number of Files Distribution (Terrorist)

<table>
<thead>
<tr>
<th>Type</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexable Files</td>
<td>80%</td>
</tr>
<tr>
<td>Multimedia Files</td>
<td>16%</td>
</tr>
<tr>
<td>Archive Files</td>
<td>4%</td>
</tr>
<tr>
<td>Non-Standard Files</td>
<td>4%</td>
</tr>
</tbody>
</table>

Volume Distribution (Terrorist)

<table>
<thead>
<tr>
<th>Type</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexable Files</td>
<td>48%</td>
</tr>
<tr>
<td>Multimedia Files</td>
<td>4%</td>
</tr>
<tr>
<td>Archive Files</td>
<td>39%</td>
</tr>
<tr>
<td>Non-Standard Files</td>
<td>9%</td>
</tr>
</tbody>
</table>
Forum Identification
-- Overall Distribution by ISP Providers

![Diagram showing distribution of forums by ISP providers for different regions:
- Middle-Eastern: Websites = 48, Yahoo! Groups = 20, Google Groups = 0, MSN = 0, AOL = 0, Local ISP = 0
- Latin-American: Websites = 4, Yahoo! Groups = 11, Google Groups = 32, MSN = 5, AOL = 0, Local ISP = 8
- US Domestic: Websites = 18, Yahoo! Groups = 31, Google Groups = 47, MSN = 9, AOL = 5, Local ISP = 0]
System Design

Forum Identification
- Identification
  - Identify extremist groups
  - Identify forums from websites
  - Identify forums from public ISPs

Forum Preprocessing
- Accessibility
  - Apply for memberships
  - Identify spidering parameters
  - Identify proper proxies

- Structure
  - Identify site maps
  - URL Ordering Features
  - URL Ordering Techniques

Wrapper Generation

Forum Storage and Analysis
- Recall Improvement
  - Log files
- Duplicate Multimedia Removal
  - Filtered collection
- Data Backup and Storage
- Statistics Generation

Forum Collection
Step 3: Identify appropriate proxies

We use three types of proxy servers

- **Transparent** proxy servers: they provide anyone with your real IP address.

- **Translucent** proxy servers: they hide your IP address or modify it in some way to prevent the target server from knowing about it. But they let anyone know that you are surfing through a proxy server.

- **Opaque** proxy servers (**preferred**): they hide your IP address and do not even let anyone know that you are surfing through a proxy server.
Step 1: Identify site maps

- Forums typically have hierarchical structures with boards, threads, and messages (Yih et al., 2004; Glance et al., 2005).
- They also contain considerable additional information.
  - Such as message posting interfaces, search, and calendar pages.
- We identify the site map of the forum based on the forum software packages.
  - Glance et al. (2005) noted that although there are only a handful of commonly used forum software packages, they are highly customizable.

<table>
<thead>
<tr>
<th>List of Forum Software</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Software</strong></td>
</tr>
<tr>
<td>Crosstar</td>
</tr>
<tr>
<td>DCForum</td>
</tr>
<tr>
<td>ezboard</td>
</tr>
<tr>
<td>IM</td>
</tr>
<tr>
<td>Invision</td>
</tr>
<tr>
<td>newbb</td>
</tr>
<tr>
<td>phpBB</td>
</tr>
<tr>
<td>rafia</td>
</tr>
<tr>
<td>vBulletin</td>
</tr>
<tr>
<td>WebRing</td>
</tr>
<tr>
<td>WebWiz</td>
</tr>
<tr>
<td>YaBB</td>
</tr>
</tbody>
</table>
System Design: Dark Web Forum Crawler Interface
Mosul Suicide Bombing Video

Phase 1: Planning Session

Phase 2: Farewell Hug

Phase 3: Suicide Truck Preparation

Phase 4: Execution
link analysis ... content analysis ... web metrics analysis ... authorship analysis ... sentiment analysis ... video analysis ...
MDS Visualization of Arab Group Web Sites

Hizb-UT-Tahrir

Hizballah Cluster

Palestinian terrorists

Jihad Supporters

Palestinian supporters
Comparison - Content Analysis

U.S. Domestic Terrorist Web sites

Middle Eastern Terrorist Web sites
Technical Sophistication

- Technical sophistication (TS)
  - To study the level of advancement of the techniques used by terrorists to establish and maintain their Web presence.
  - Table 1 shows the TS Measures identified from Palmer and David (1998).

<table>
<thead>
<tr>
<th>Measures</th>
<th>Weights</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic HTML Techniques</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of Lists</td>
<td>0/1</td>
<td>All attributes can be automatically identified from terrorist Websites using programs.</td>
</tr>
<tr>
<td>Use of tables</td>
<td>0/2</td>
<td></td>
</tr>
<tr>
<td>Use of Frames</td>
<td>0/2</td>
<td></td>
</tr>
<tr>
<td>Use of Forms</td>
<td>0/1.5</td>
<td></td>
</tr>
<tr>
<td>Embedded Multimedia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of Background Image</td>
<td>0/1</td>
<td></td>
</tr>
<tr>
<td>Use of Background Music</td>
<td>0/2</td>
<td></td>
</tr>
<tr>
<td>Stream Audio/Video</td>
<td>0/3.5</td>
<td></td>
</tr>
<tr>
<td>Advanced HTML</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of phtml/shtml</td>
<td>0/2.5</td>
<td></td>
</tr>
<tr>
<td>Use Predefined Functions?</td>
<td>0/2</td>
<td></td>
</tr>
<tr>
<td>Use Self-defined Functions?</td>
<td>0/4.5</td>
<td></td>
</tr>
<tr>
<td>Dynamic Web Programming</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use CGI</td>
<td>0/2.5</td>
<td></td>
</tr>
<tr>
<td>Use PHP</td>
<td>0/4.5</td>
<td></td>
</tr>
<tr>
<td>Use JSP/ASP</td>
<td>0/5.5</td>
<td></td>
</tr>
</tbody>
</table>
Analysis Results: Technical Sophistication

- Overall, the technical sophistication of terrorist Web sites is on par with US government Web sites.

- US government Web sites are better at the use of basic HTML techniques and dynamic Web programming.

- Terrorist Web sites are using more embedded multimedia.

Table 4. Technical Sophistication Comparison Results

<table>
<thead>
<tr>
<th>High-level Attributes</th>
<th>Weighted Average Score</th>
<th>t-Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>US</td>
<td>Terrorists</td>
</tr>
<tr>
<td>Basic HTML Techniques</td>
<td>0.913043</td>
<td>0.710526</td>
</tr>
<tr>
<td>Embedded Multimedia</td>
<td>0.565217</td>
<td>0.833333</td>
</tr>
<tr>
<td>Dynamic HTML</td>
<td>1.789855</td>
<td>1.771929</td>
</tr>
<tr>
<td>Dynamic Web Programming</td>
<td>2.159420</td>
<td>1.407894</td>
</tr>
<tr>
<td>Average</td>
<td>1.356884</td>
<td>1.180921</td>
</tr>
</tbody>
</table>
Overall, the Middle Eastern terrorist groups’ Web sites display a high level of technical sophistication that is on par with the U.S. government Web sites.

- Terrorist Web sites intensively use embedded media

Flash animations and pictures depicting Marxist symbols and personalities on the Iranian People’s Fadaee Guerilla web site
Analysis Results: Web Interactivity

- At Web interactivity level, terrorist Web sites do not show significant differences from US government Web sites.

- However, terrorist Web sites are doing much better in supporting community-based interaction by providing online forums and chat rooms; while few government Web sites do.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Weighted Average Score</th>
<th>t-Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-to-one</td>
<td>0.342857</td>
<td>0.292169</td>
</tr>
<tr>
<td>Community</td>
<td>0.028571</td>
<td>0.168675</td>
</tr>
<tr>
<td>Transaction</td>
<td>0.3</td>
<td>Not presented</td>
</tr>
<tr>
<td>Average</td>
<td>0.185714</td>
<td>0.230422</td>
</tr>
</tbody>
</table>

Table 6. Web Interactivity Comparison Results
Sentiment Analysis: Research Design

- Proposed Affect Analysis Approach
  1. Create Hate and Violence Affect Lexicons
  2. Measure Affect Intensity using score based approach
  3. Overcome affect ambiguity using Probabilistic Disambiguation
  4. Measure and visualize Affect Intensity Variation using Principal Component Analysis
    - Message Level Variation
    - Temporal Variation
AVM Algorithm Illustration

**AVM-M**
- **Messages**
  - 1
  - 2
  - 3
- **Feature Vectors**
  - $x_1$
  - $x_2$
  - $x_3$
- **Dimensionality Reduction**
  - $e_1 = a_1^T x_i$
  - $e_2 = a_2^T x_i$
- **2D Projection**
- **Affect Variation**
  - $v = \frac{\sum_{i=1}^{n} d_i}{n}$

**AVM-T**
- **Sliding Window**
  - $j$
  - $w$
- **Messages**
  - 1
  - 2
  - 3
- **Feature Vectors**
  - $x_1$
  - $x_2$
- **Dimensionality Reduction**
  - $e_1 = a_1^T x_i$
  - $e_2 = a_2^T x_i$
- **2D Projection**
- **Affect Variation**
  - $v = \frac{\sum_{i=1}^{n} d_i}{n}$
6. Evaluation

- The dataset consists of messages from Domestic and Middle Eastern forums.

<table>
<thead>
<tr>
<th>Forum</th>
<th>Authors</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angelic Adolf</td>
<td>28</td>
<td>78</td>
</tr>
<tr>
<td>Aryan Nation</td>
<td>54</td>
<td>189</td>
</tr>
<tr>
<td>CCNU</td>
<td>2</td>
<td>229</td>
</tr>
<tr>
<td>Neo-Nazi</td>
<td>98</td>
<td>632</td>
</tr>
<tr>
<td>NSM</td>
<td>289</td>
<td>2654</td>
</tr>
<tr>
<td>Smash Nazi</td>
<td>10</td>
<td>66</td>
</tr>
<tr>
<td>White Knights</td>
<td>24</td>
<td>751</td>
</tr>
<tr>
<td>World Knights</td>
<td>35</td>
<td>77</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>540</strong></td>
<td><strong>4676</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Forum</th>
<th>Authors</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azzamy</td>
<td>60</td>
<td>337</td>
</tr>
<tr>
<td>Friends in Islam</td>
<td>119</td>
<td>339</td>
</tr>
<tr>
<td>Islamic Union</td>
<td>67</td>
<td>473</td>
</tr>
<tr>
<td>Kataeb</td>
<td>66</td>
<td>229</td>
</tr>
<tr>
<td>Kataeb Qassam</td>
<td>57</td>
<td>582</td>
</tr>
<tr>
<td>Taybah</td>
<td>63</td>
<td>290</td>
</tr>
<tr>
<td>Osama Lover</td>
<td>42</td>
<td>173</td>
</tr>
<tr>
<td>Wa Islamah</td>
<td>363</td>
<td>926</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>837</strong></td>
<td><strong>3349</strong></td>
</tr>
</tbody>
</table>
7. Results: Intensity Relationship

Affect Regression Analysis: Message Level

<table>
<thead>
<tr>
<th></th>
<th>U.S.</th>
<th>Middle Eastern</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>4676</td>
<td>3349</td>
</tr>
<tr>
<td>beta (slope)</td>
<td>0.079</td>
<td>0.682</td>
</tr>
<tr>
<td>t-Stat</td>
<td>21.354</td>
<td>48.265</td>
</tr>
<tr>
<td>P-Value</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>R-Square</td>
<td>0.076</td>
<td>0.486</td>
</tr>
</tbody>
</table>

Strong hate and violence Correlation, especially for Middle-Eastern group.
7. Results: Intensity Relationship

Affect Regression Analysis: Forum Level

<table>
<thead>
<tr>
<th></th>
<th>U.S.</th>
<th>Middle Eastern</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>beta (slope)</td>
<td>0.347</td>
<td>0.471</td>
</tr>
<tr>
<td>t-Stat</td>
<td>1.760</td>
<td>10.306</td>
</tr>
<tr>
<td>P-Value</td>
<td>0.139</td>
<td>0.000</td>
</tr>
<tr>
<td>R-Square</td>
<td>0.383</td>
<td>0.947</td>
</tr>
</tbody>
</table>
1. Incoming Message

Arabic Feature Extraction Component

2. Elongation Filter

Filtered Message

3. Root Dictionary

Root Clustering Algorithm

4. Similarity Scores (SC)

max(SC) + 1

Feature Set

Generic Feature Extractor
1. Compute eigenvectors for 2 principal components of feature group

\[ \Phi_x, \Phi_y \]

| 0.533 | 0.956 |
| -0.541 | 0.445 |
| 0.034 | 0.089 |
| 0.653 | 0.456 |
| 0.975 | -0.085 |
| 0.143 | -0.381 |

2. Extract feature usage vectors

\[ 1,0,0,2,1,2 \]

3. Transform into 2-dimensional space

\[ x = \sum Z\Phi_x \]
\[ y = \sum Z\Phi_y \]

Repeat steps 2 and 3

Message Text

WANTED: MS OEM & RETAIL SOFTWARE-NETWORK CARDS-SUN & CISCO

We are buying Cisco and Sun Hardware as well as Microsoft OEM & Retail Software.
We are also looking to buy used and new network cards.
PLEASE LOOK BELOW...

We are buying MS WIN 95, 98, 98SE, 2K, XP, OFFICE 2000, XP, 97, 95 PRO-SEE & MORE...

Looks at the list below.

We are also buying covers & manuals with COA's.
WE ARE ALSO BUYING COA's.

WTB: COVERS/ MANUALS WITH COA's WIN 95, 98, 98 SE, NT 4.0 & MORE...

Feature Usage Vector Z

\[ 0,1,3,0,1,0 \]
Special Char. Writeprints

Author A

Author B

Author C

Author D

<table>
<thead>
<tr>
<th>Feature</th>
<th>( \Phi_x )</th>
<th>( \Phi_y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>~</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>@</td>
<td>0.022814</td>
<td>-0.01491</td>
</tr>
<tr>
<td>#</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$</td>
<td>-0.01253</td>
<td>-0.17084</td>
</tr>
<tr>
<td>%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>^</td>
<td>-0.01227</td>
<td>-0.01744</td>
</tr>
<tr>
<td>&amp;</td>
<td>-0.01753</td>
<td>-0.0777</td>
</tr>
<tr>
<td>*</td>
<td>-0.03017</td>
<td>-0.05931</td>
</tr>
<tr>
<td>-</td>
<td>-0.12656</td>
<td>0.991784</td>
</tr>
<tr>
<td>_</td>
<td>0.998869</td>
<td>0.047184</td>
</tr>
<tr>
<td>=</td>
<td>-0.05113</td>
<td>-0.07576</td>
</tr>
<tr>
<td>+</td>
<td>0.142534</td>
<td>0.021726</td>
</tr>
<tr>
<td>&gt;</td>
<td>-0.1077</td>
<td>0.392182</td>
</tr>
<tr>
<td>&lt;</td>
<td>-0.10618</td>
<td>0.213193</td>
</tr>
<tr>
<td>[</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>]</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>{</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>}</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>/</td>
<td>-0.05075</td>
<td>-0.09065</td>
</tr>
<tr>
<td>\</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.05965</td>
</tr>
</tbody>
</table>
## Experiment Results

<table>
<thead>
<tr>
<th>Features</th>
<th>English Dataset</th>
<th>Arabic Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C4.5</td>
<td>SVM</td>
</tr>
<tr>
<td>F1</td>
<td>85.76%</td>
<td>88.0%</td>
</tr>
<tr>
<td>F1+F2</td>
<td>87.23%</td>
<td>90.77%</td>
</tr>
<tr>
<td>F1+F2+F3</td>
<td>88.30%</td>
<td>96.5%</td>
</tr>
<tr>
<td>F1+F2+F3+F4</td>
<td>89.10%</td>
<td>87.90%</td>
</tr>
</tbody>
</table>

![Graph showing comparison of English and Arabic datasets](image-url)
ClearGuidance.com.com Analysis using CyberGate
CyberGate Design Framework

- CyberGate is designed to analyze computer mediated communication content.
  - It includes a rich feature set, and multiple feature selection and visualization methods.
System Design: CyberGate System Design

Information Types
- Ideational
  - Affect
  - Sentiment
  - Topic
- Textual
  - Genre
  - Style

Feature Set
- Language Resources
  - Thesaurus
  - Lexicons
  - Lexical
  - Syntactic
  - Structural
- Process Resources
  - Part-of-Speech Definitions
  - N-gram Definitions
  - Statistical Feature Definitions

Feature Extractor
- CMC Text
  - Input text
- Reference
  - Get feature list
- Feature Vectors

Feature Reduction
- Projection
  - Principal Component Analysis
  - Multi-Dimensional Scaling
- Ranking
  - Information Gain
  - Decision Tree Models

Visualization
- Multi-Dimensional
  - Writeprints
  - Parallel Coordinates
  - Radar Charts
  - MDS Plots

Text Overlay
- Ink Blots
- Text Annotation
<table>
<thead>
<tr>
<th>Resource</th>
<th>Category</th>
<th>Feature Groups</th>
<th>Quantity</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Lexical</td>
<td>Word Length</td>
<td>20</td>
<td>word frequency distribution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Letters</td>
<td>26</td>
<td>A,B,C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Special Characters</td>
<td>21</td>
<td>$,@,#,*,&amp;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Digits</td>
<td>10</td>
<td>0,1,2</td>
</tr>
<tr>
<td>Syntactic</td>
<td></td>
<td>Function Words</td>
<td>250</td>
<td>of, for, the, on, if</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pronouns</td>
<td>20</td>
<td>I, he, we, us, them</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Conjunctions</td>
<td>30</td>
<td>and, or, although</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prepositions</td>
<td>30</td>
<td>at, from, onto, with</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Punctuation</td>
<td>8</td>
<td>!,?,;:&quot;</td>
</tr>
<tr>
<td>Structural</td>
<td></td>
<td>Document Structure</td>
<td>14</td>
<td>has greeting, has url, requoted content</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technical Structure</td>
<td>50</td>
<td>file extensions, fonts, images</td>
</tr>
<tr>
<td>Lexicons</td>
<td></td>
<td>Sentiment Lexicons</td>
<td>3000</td>
<td>positive, negative terms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Affect Lexicons</td>
<td>5000</td>
<td>happiness, anger, hate, excitement</td>
</tr>
<tr>
<td>Process</td>
<td>Lexical</td>
<td>Word-Level Lexical</td>
<td>8</td>
<td>% char per word</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Char-Level Lexical</td>
<td>7</td>
<td>% numeric char per message</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vocabulary Richness</td>
<td>8</td>
<td>hapax legomana, Yules K,</td>
</tr>
<tr>
<td>Syntactic</td>
<td></td>
<td>POS Tags</td>
<td>2200</td>
<td>NP_VB</td>
</tr>
<tr>
<td>Content-Based</td>
<td></td>
<td>Noun Phrases</td>
<td>Varies</td>
<td>account, bonds, stocks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Named Entities</td>
<td>Varies</td>
<td>Enron, Cisco, El Paso, California</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bag-of-words</td>
<td>Varies</td>
<td>all words except function words</td>
</tr>
<tr>
<td>N-Grams</td>
<td></td>
<td>Character-Level</td>
<td>Varies</td>
<td>aa, ab, aaa, aab</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Word-Level</td>
<td>Varies</td>
<td>went to, to the, went to the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>POS-Level</td>
<td>Varies</td>
<td>NNP_VB VB VB ADJ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Digit Level</td>
<td>1100</td>
<td>12, 94, 192</td>
</tr>
</tbody>
</table>
CyberGate: Multi-Dimensional Views

Two dimensional PCA projections based on feature occurrences. Each circle denotes a single message. Selected message is highlighted in pink. Writeprints show feature usage/occurrence variation patterns. Greater variation results in more sporadic patterns.

Writeprints

Parallel vertical lines represent features. Bolded numbers are feature numbers (0-15). Smaller numbers above and below feature lines denote feature range. Blue polygonal lines represent messages. Selected message is highlighted in red. Selected feature is highlighted in pink (#2).

Parallel Coordinates

Chart shows normalized feature usage frequencies. Blue line represents author’s average usage, red line indicates mean usage across all authors, and green line is another author (being compared against). The numbers represent feature numbers. Selected feature is highlighted (#6).

Radar Charts

MDS algorithm used to project features into two-dimensional space based on occurrence similarity. Each circle denotes a feature. Closer features have higher co-occurrence. Labels represent feature descriptions. Selected feature is highlighted in pink (the term “services”).

MDS Plots
Feature occurrences are highlighted in blue. The selected bag-of-words feature is highlighted in red ("CounselEnron").

Colored circles (blots) superimposed onto feature occurrence locations in text. Blot size and color indicates feature importance and usage. Selected feature's blots are highlighted with black circles.
ClearGuidance.com

• Background
  – Forum with some members affiliated with Toronto terror plot.
  – Reportedly had as many as 15,000 members.
  – Unfortunately, the site went offline in February of 2004, before we began spidering forums.
  – We were able to retrieve selected content from various blogs.

| Authors | Messages | Duration        |
|---------|----------|-----------------
• Member locations
  – Shown to the right are the self-reported member locations.
  – Approximately 2/3 of the 269 members specified a location country.
  – Breakdowns for those that did specify:
    • Majority located in USA, UK, Canada, and the Middle East.
• Member Interaction Network
  – Constructed using the Hybrid Interaction Coherence (HIC) Algorithm
  – Blue nodes indicate members with the greatest number of in-links.
  – These members are the core set of forum “experts” and propagandists
Forum Members

Writeprint view showing bag-of-word feature usage variation
Uses a sliding window with each point representing a text window
The pattern suggests that his author has considerable topical variation, discussing several different things.

MDS projections of bag-of-word features.
Features are grouped based on occurrence similarity.
Topics include BBC, Robert Kilroy, and Chechnya.

Ink Blots (colored circles) superimposed onto feature occurrence locations in text.
Blot size and color indicates feature importance and usage.
Author has few key features.
The series of overlapping circular patterns for bag-of-word features indicates that the author’s discussion revolves around a related set of topics.

Bag-of-words are predominantly related to religious topics.

Many large red blots indicative of the presence of features unique to this author.

This author attempts to use his religious “expertise”.

same thing, but even if I sat down during it nobody would have the guts too say anything to me. My school is full of wussies and wannabe "gangstas". _______________ Say (t the Rejecters): "My Lord is not uneasy because of you if ye call not on Him: But ye have indeed rejected (Him), and soon will come the inevitable (punishment)!” Al-Furqan Vol 4, Book 56. Prophets. Hadith 543. Narrated by Abu Huraira: The Prophet said, "Allah created Adam, making him 60 cubits tall. When He created him, He said to him, "Go and greet that group of angels, and listen to their reply, for it will be your greeting (salutation) and the greeting (salutations of your offspring." So, Adam said (to the angels), "As-Salamu Alaikum (i.e. Peace be upon you)." The angels said, "As-salamu Alaika wa Rahmatu-l-Lahi " (i.e. Peace and Allah's#Mercy be upon you). Thus the angels added to Adam’s salutation the expression, "$\#\$ Wa Rahmatu-l-Lahi, $\#\$ Any person who will speak Arabic will recognize Allah’s bounty. " 
This author also has a circular bag-of-word variation pattern reflective of topical consistency.

Radar chart showing bag-of-word feature usages (each feature numbered).
Comparison of author usage to forum mean.
Selected feature is use of term “jihaad” which is far higher than the mean.

Text annotation view showing key bag-of-words highlighted.
Selected feature (i.e., “jihaad”) is shown in red.
This author constantly attempts to glorify acts of violence and suicide bombings.
Impact of Improper Guidance

This author was later arrested as a major culprit in the Toronto terror plot ("Soldier of God"). He uses many violent affect terms.

Radar chart showing violent affect feature usages.

Comparison to mean shows several high occurrence terms (e.g., jihad, martyrdom).

Selected feature is use of term "jihad" which is the highest in the forum.

Text annotation view showing key bag-of-words highlighted.

Selected feature (i.e., "jihad") is shown in red.

This author constantly attempts to justify acts of violence and terrorism.

“…there are so many paid sheikhs stuck in this life….no point going to them for fatwas…personally speaking…cuz they don’t even agree with jihad in the first place”
Case Study: Author Interaction Network across Forums

World Knights
White Knights
Angellic Adolf
Aryan Nation
Smash Nazi
Neo-Nazi
CCNU
World Knights
White Knights
Angellic Adolf
Aryan Nation
Smash Nazi
Neo-Nazi
CCNU
NSM World
IEDs in the Dark Web
Specialized Content Procedure
The lexicon is composed of over 100 relevant terms and their variations (inflective forms).

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Variations</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>المدى الفعّال</td>
<td>المدى الفعال</td>
<td>Effective Range</td>
</tr>
<tr>
<td>ألمدى ألفعال</td>
<td>ألمدى ألفعال</td>
<td></td>
</tr>
<tr>
<td>ألمدى ألفعال</td>
<td>ألمدى ألفعال</td>
<td></td>
</tr>
<tr>
<td>ألمدى ألفعال</td>
<td>ألمدى ألفعال</td>
<td></td>
</tr>
<tr>
<td>العبوات</td>
<td>العبوات</td>
<td>Explosives</td>
</tr>
<tr>
<td>الأسلحة</td>
<td>الأسلحة</td>
<td>Weapons, “The Weapons”</td>
</tr>
<tr>
<td>أسلحة</td>
<td>أسلحة</td>
<td></td>
</tr>
<tr>
<td>أسلحة</td>
<td>أسلحة</td>
<td></td>
</tr>
<tr>
<td>التصنيع الكيماوي</td>
<td>التصنيع الكيماوي</td>
<td>Chemical Manufacturing</td>
</tr>
<tr>
<td>تصنيع الكيماوي</td>
<td>تصنيع الكيماوي</td>
<td></td>
</tr>
</tbody>
</table>
Using the lexicon, we used a search engine to extract all web pages with these terms from our collection.

- A total of 2541 relevant web pages were collected from 30 web sites.
- Over 90% of these pages came from a core set of 7 web sites.

A breakdown of the collected pages is presented below:

<table>
<thead>
<tr>
<th>Total Web Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Web Sites</td>
</tr>
<tr>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Core Web Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Site</td>
</tr>
<tr>
<td><a href="http://www.qudsay.com">www.qudsay.com</a></td>
</tr>
<tr>
<td><a href="http://www.albasrah.net">www.albasrah.net</a></td>
</tr>
<tr>
<td><a href="http://www.khayma.com">www.khayma.com</a></td>
</tr>
<tr>
<td><a href="http://www.jamaat.org">www.jamaat.org</a></td>
</tr>
<tr>
<td><a href="http://www.hilafet.com">www.hilafet.com</a></td>
</tr>
<tr>
<td><a href="http://www.geocities.com">www.geocities.com</a></td>
</tr>
</tbody>
</table>
Segmentation: Manual Inspection

• The retrieved pages were manually inspected by a domain expert familiar with the Middle Eastern Dark Web collection.

• The expert manually classified the web pages into two categories:
  – Discussion Pages
  – Materials Pages

<table>
<thead>
<tr>
<th>Content Type</th>
<th>No. Web Sites</th>
<th>No. Web Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussion</td>
<td>30</td>
<td>2501</td>
</tr>
<tr>
<td>Materials</td>
<td>4</td>
<td>40</td>
</tr>
</tbody>
</table>
2 Attack on a military patrol that was lured to a minefield north of the city of Tikrit. Two American soldiers were killed, four others were injured and three vehicles were destroyed.
Explosive materials
1 - Trinitrotoluene TNT:

Composition

Preparation Procedure

Source:
http://66.148.85.35/site/books/dawra-mtfajerat.zip/dawra-mtfajerat.doc
## Extended Arabic Feature Set

<table>
<thead>
<tr>
<th>Group</th>
<th>Category</th>
<th>Quantity</th>
<th>Description/Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical</strong></td>
<td>Word-Level</td>
<td>5</td>
<td>total words, % char. per word</td>
</tr>
<tr>
<td></td>
<td>Character-Level</td>
<td>5</td>
<td>total char., % char. per message</td>
</tr>
<tr>
<td></td>
<td>Character N-Grams</td>
<td>&lt; 18,278</td>
<td>count of letters, char. bigrams, trigrams (e.g., کک،اف)</td>
</tr>
<tr>
<td></td>
<td>Digit N-Grams</td>
<td>&lt; 1,110</td>
<td>count of digits, digit bigrams, digit trigrams (e.g., 1, 12, 123)</td>
</tr>
<tr>
<td>Word Length Distribution</td>
<td></td>
<td>20</td>
<td>frequency distribution of 1-20 letter words</td>
</tr>
<tr>
<td>Vocabulary Richness</td>
<td></td>
<td>8</td>
<td>richness (e.g., hapax legomena, Yule’s K, Honore’s H)</td>
</tr>
<tr>
<td>Special Characters</td>
<td></td>
<td>21</td>
<td>occurrences of special char. (e.g., @#$%^&amp;*+=)</td>
</tr>
<tr>
<td><strong>Syntactic</strong></td>
<td>Function Words</td>
<td>300</td>
<td>frequency distribution of function words (e.g., of, for, to)</td>
</tr>
<tr>
<td></td>
<td>Punctuation</td>
<td>12</td>
<td>occurrence of punctuation marks (e.g., !;:,?)</td>
</tr>
<tr>
<td></td>
<td>Word Root N-Grams</td>
<td>varies</td>
<td>roots, bigrams, trigrams (e.g., كسب،كتب)</td>
</tr>
<tr>
<td><strong>Structural/HTML</strong></td>
<td>Message-Level</td>
<td>6</td>
<td>e.g., has greeting, has url, requoted content</td>
</tr>
<tr>
<td></td>
<td>Paragraph-Level</td>
<td>8</td>
<td>e.g., number of paragraphs, sentences per paragraph</td>
</tr>
<tr>
<td></td>
<td>Technical Structure</td>
<td>50</td>
<td>e.g., file extensions, fonts, use of images, HTML tags</td>
</tr>
<tr>
<td></td>
<td>HTML Tag N-Grams</td>
<td>&lt; 46,656</td>
<td>e.g., &lt;head&gt;, &lt;br&gt;, &lt;td&gt;, &lt;message&gt;</td>
</tr>
<tr>
<td><strong>Topical</strong></td>
<td>Word N-Grams</td>
<td>varies</td>
<td>bag-of-words n-grams (e.g., “explosive”, “explosive device”)</td>
</tr>
</tbody>
</table>
Categorization

- The table and graph summarize the 100 bootstrapping instance results.

- Using feature selection, we were able to get 88.8% accuracy.

- We were also able to isolate a subset of approximately 9,000 key features.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Features</th>
<th>Mean Accuracy</th>
<th>Standard Deviation</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>21,333</td>
<td>81.938</td>
<td>5.313</td>
<td>65.00 – 92.50</td>
</tr>
<tr>
<td>SVM-IG</td>
<td>9,268</td>
<td>88.838</td>
<td>3.238</td>
<td>80.00 – 96.25</td>
</tr>
</tbody>
</table>

Classification Results

![Classification Results graph]

- **SVM**
- **SVM-IG**
We used link analysis to construct the web site network for the Middle Eastern groups in our Dark Web collection.

- The network includes Dark Web groups as well as sites linked to by these groups (“other” sites).
- Only sites with some linkage to others (in or out links) were included.

The nodes were color coded based on their content:

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>No. of Sites</th>
<th>No. Linked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Materials</td>
<td>Containing material pages</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Discussion</td>
<td>Containing discussion pages</td>
<td>30</td>
<td>19</td>
</tr>
<tr>
<td>Dark Web</td>
<td>Additional Dark Web Sites</td>
<td>120</td>
<td>45</td>
</tr>
<tr>
<td>Other Sites</td>
<td>Linked to by the Dark Web Sites</td>
<td>1,671</td>
<td>192</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>All sites linking to or from Dark Web sites</strong></td>
<td><strong>1,825</strong></td>
<td><strong>258</strong></td>
</tr>
</tbody>
</table>
Link Analysis: Dark Web Network

www.albasrah.net
Major Iraqi resistance web site

www.saad.net
Major Dark Web site

www.geocities.com/maoso3ah
Training materials

www.geocities.com/m_ale3dad4
Training materials
www.albasrah.net
Is a website with links to the former Iraqi Baathist regime. It contains a large collection of war images and reports of military operations by Iraqi insurgents.

www.geocities.com/m_ale3dad4
Is a collection of training material. Topics include weapons, their usage, and the manufacturing of IEDs. The website contains video demonstrations, books, and other documents in English and Arabic.

www.geocities.com/maoso3ah
Is an “encyclopedia” of military training and preparation for Jihad.

www.saaid.net
Is an Islamic directory. Much of the content pertains to unrelated topics. However, some of the contributors support the Jihadi Salafi movement.
Segmentation: Dissemination of Videos in Online Discussion Forums

Videos Produced by Iraqi Insurgency Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al-Jaysh al-Islami fil-'Iraq</td>
<td><img src="image1.jpg" alt="Image" /></td>
</tr>
<tr>
<td>(the Islamic Army in Iraq)</td>
<td><img src="image2.jpg" alt="Image" /></td>
</tr>
<tr>
<td>Tandhim al-Qa'ida fi Bilad al-Rafidayn</td>
<td><img src="image3.jpg" alt="Image" /></td>
</tr>
<tr>
<td>(al-Qaeda's Organisation in Mesopotamia)</td>
<td><img src="image4.jpg" alt="Image" /></td>
</tr>
<tr>
<td>Jaysh Ansar al-Sunna</td>
<td><img src="image5.jpg" alt="Image" /></td>
</tr>
<tr>
<td>(Partisans of the Sunna Army)</td>
<td><img src="image6.jpg" alt="Image" /></td>
</tr>
</tbody>
</table>

Re-packaging and Disseminating by Sympathizers

- al-Jabhah al-'ilamiyah al-islamiyah al-'alamiyah
- (Global Islamic Media Front)

Video is converted into different sizes and formats

<table>
<thead>
<tr>
<th>Format</th>
<th>Size (MB)</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>VCD</td>
<td>290</td>
<td>1</td>
</tr>
<tr>
<td>WMV</td>
<td>107</td>
<td>15</td>
</tr>
<tr>
<td>RMVB</td>
<td>93</td>
<td>9</td>
</tr>
<tr>
<td>WMV</td>
<td>53</td>
<td>9</td>
</tr>
<tr>
<td>RMVB</td>
<td>46</td>
<td>9</td>
</tr>
<tr>
<td>WMV</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>RMVB</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>RMVB</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>3GP</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>GIF</td>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td>TOTAL</td>
<td>673</td>
<td>82</td>
</tr>
</tbody>
</table>

“Jihad Academy” video is posted on different forums

- [http://www.la7oood.com](http://www.la7oood.com) 01/20/2006 01:53 PM
- [http://www.almarsaa.net](http://www.almarsaa.net) 01/26/2006 03:19 AM
# Segmentation: Video Files

<table>
<thead>
<tr>
<th>Classification</th>
<th># Files</th>
<th>File Size (MB)</th>
<th>PlayTime (hh:mm:ss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attack</td>
<td>313</td>
<td>2,479.26</td>
<td>39:12:59</td>
</tr>
<tr>
<td>Statement</td>
<td>129</td>
<td>1,784.15</td>
<td>45:37:51</td>
</tr>
<tr>
<td>Propaganda</td>
<td>118</td>
<td>1,153.81</td>
<td>16:58:54</td>
</tr>
<tr>
<td>Produced</td>
<td>29</td>
<td>556.35</td>
<td>07:49:54</td>
</tr>
<tr>
<td>Beheading</td>
<td>68</td>
<td>294.95</td>
<td>04:43:03</td>
</tr>
<tr>
<td>Training</td>
<td>9</td>
<td>196.49</td>
<td>03:20:12</td>
</tr>
<tr>
<td>Hostage Taking</td>
<td>25</td>
<td>128.16</td>
<td>01:43:25</td>
</tr>
<tr>
<td>Tribute</td>
<td>12</td>
<td>113.93</td>
<td>02:35:11</td>
</tr>
<tr>
<td>Instructional</td>
<td>1</td>
<td>16.72</td>
<td>00:08:24</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>704</strong></td>
<td><strong>6,723.83</strong></td>
<td><strong>122:06:53</strong></td>
</tr>
</tbody>
</table>
Interface for Multimedia Coder Tool (MCT) ...(1)

<table>
<thead>
<tr>
<th>Coder ID</th>
<th>File Name</th>
<th>File Path</th>
<th>Content Category</th>
<th>Content Feature</th>
<th>Value</th>
<th>Comments</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coder_1</td>
<td></td>
<td>\ai-arab-2\D\new sample\18.wmv</td>
<td>Production</td>
<td>Planning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Extremist</td>
<td>Message</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Organization</td>
<td>Threats</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Video
<table>
<thead>
<tr>
<th>Class</th>
<th>Content Category</th>
<th>Content Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Violent Attacks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attack</td>
<td>Planning</td>
<td>All kinds of attacks except hostage taking and suicide bombing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statement</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Target</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weapon used (e.g., bombing)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hostage Taking</td>
<td>Captive’s name</td>
<td>Person held against his/her will</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Captive’s nationality</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Demand</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Execution</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hostage Negotiation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suicide Bombing</td>
<td>Method</td>
<td>Bombing carried out by a person who does not hope to survive it</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Suicider’s name</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Suicider’s nationality</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statement</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Target</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Message</td>
<td>Tribute</td>
<td>Statement by prominent members of the extremist group [24]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leader Statement</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Newsletter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>Instruction</td>
<td>Instruction is for teaching specific skills, while training is for documenting activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>Content Category</td>
<td>Content Feature</td>
<td>Description</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>2. Content Properties</td>
<td>Victim</td>
<td>Number of person(s)</td>
<td>Injury or loss of life</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Name(s)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nationality</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Civilian/Military</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date of Attack</td>
<td>Hijri (Islamic) Calendar</td>
<td>Reported attack date in the video</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Julian Calendar</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>City</td>
<td>Location of depicted event</td>
</tr>
<tr>
<td></td>
<td></td>
<td>County</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>District</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Town</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Country</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Target of threat</td>
<td>Person</td>
<td>Threatened entity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Organization</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Country</td>
<td></td>
</tr>
<tr>
<td>3. Groups</td>
<td>Extremist group</td>
<td>Reported name</td>
<td>Group involved</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sub-group</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Media agency</td>
<td></td>
</tr>
<tr>
<td>4. Expressions</td>
<td>Quotation</td>
<td>Religious quotation</td>
<td>Quotation used</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Violent quotation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reference to Media</td>
<td>Arab Media</td>
<td>Mention of media</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Western Media</td>
<td></td>
</tr>
<tr>
<td>5. Production</td>
<td>Production Feature</td>
<td>Title of the video</td>
<td>Video feature</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sound</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Visual (e.g., subtitle)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>URL</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiclip</td>
<td></td>
</tr>
</tbody>
</table>
Segmentation: Implanting and setting off the device

- Video 28: Insurgents prepare the IED
- Location: Mushahad region, Iraq
- Insurgents: Islamic Front of the Iraqi Resistance

Planting the device  Hiding the device  Setting it off
Segmentation: Targeting Pedestrian Soldiers

- Video 116: IED targeting pedestrian soldiers
- Location: Unknown
- Insurgents: Islamic Army in Iraq

Soldiers gather → Explosive is set off → Rescue
Segmentation: IED in Parked Car

- Video 198: Explosive device planted in roadside parked car
- Location: Unknown
- Insurgents: Islamic Movement, Iraqi Mujahedeen

Targeted humvee    Car with IED    Humvee approaching car    IED set off
Dark Web Research Plan

• Expanded spidering of international home-grown groups (different languages, different regions, including US)
• Dark Web portal system development (with retrieval, analysis, visualization, and machine translation abilities)
• Deep sentiment, affect, interaction, temporal and path analysis of “infectious ideas”
• Advanced content-based multimedia analysis and coding (e.g., images, videos)
• Advanced stenography analysis of encrypted messages/images
Search Engines ➔ Web Mining

- Google Anatomy and Google Story
- Inside Internet Search Engines (Excite Story)
- Vertical and Multilingual Portals: HelpfulMed and CMedPort
- Web Mining: Using Google, EBay, and Amazon APIs
- The Dark Web
For more information:

Artificial Intelligence Lab, University of Arizona

Hsinchun Chen …
hchen@eller.arizona.edu …
http://ai.arizona.edu …