



Web Mining: An Overview

Mandar Mitra

CVPR Unit

Indian Statistical Institute

Kolkata

- What is Web mining?
- Classification of Web mining tasks
- Challenges
- Web content mining
- Web structure mining
- Web usage mining
- References

What is Web Mining?

- Web mining is the automatic discovery and extraction of potentially useful and previously unknown information from Web data
- Old wine in a new bottle?
 - Web mining = databases + information retrieval + artificial intelligence (natural language processing, machine learning) + ...
- So, why the interest?
 - multidisciplinary nature
 - growth of Web information sources
 - e-commerce potential: ***“Electronic commerce is emerging as the killer domain for data-mining technology”***

- **Content mining:** mine the content of documents/pages
 - retrieval, clustering of search results, filtering, summarization, classification / categorization, etc.
- **Structure mining:** study the link structure of pages and sites
 - authorities and hubs, page ranking (Google), detection of communities
- **Usage mining:** analyze usage data, surfing behaviour/patterns
 - site restructuring, marketing
- Compartments are not water-tight
 - searching, filtering (content-based / collaborative / reputation-based)

- Unstructured and heterogeneous
- Multimedia
- Size + rapid growth
 - 1 new server every 2 hours
 - 5 million documents in 1995 to 320 million documents in 1998
- Dynamic
- Networked/distributed

- Types of data: text, images, audio, video, databases
- Text is most important
 - Unstructured – free text
 - Semi-structured – HTML documents
 - Structured – tables, documents generated from databases

Text Mining: Outline

- Indexing
- Searching
- Filtering
- Word relationships
- Classification
- Discovering document relationships
- Summarization

Text Mining: Indexing

- Any text item (“document”) represented as list of terms and associated weights

$$D = (\langle t_1, w_1 \rangle, \dots, \langle t_n, w_n \rangle)$$

- Term = keywords or content-descriptors
- Weight = measure of the importance of a term in representing the information contained in the document

- Tokenize: identify individual words
- Stopword removal: eliminate common words, e.g. and, of, the, etc.
- Stemming: reduce words to a common root
 - e.g. analysis, analyze, analyzing → analy
 - use standard algorithms (Porter)
- Thesaurus: find synonyms for words in the document
- Phrases: find multi-word terms e.g. computer science, data mining
 - use syntax/linguistic methods or “statistical” methods

Indexing: Term Weights

- Term frequency (tf): repeated words are strongly related to content
- Inverse document frequency (idf): uncommon term is more important
- Normalization by document length
 - long docs. contain many distinct words
 - long docs. contain same word many times
 - term-weights for long documents should be reduced
 - use # bytes, # distinct words, Euclidean length, etc.
- $\text{Weight} = \text{tf} \times \text{idf} / \text{normalization}$

- Measure vocabulary overlap between user query and documents

$$Sim(Q, D) = \sum_i wt(q_i) \times wt(d_i)$$

- Use inverted list (index)

$$Term_i \rightarrow (D_{i_1}, w_{i_1}), \dots, (D_{i_k}, w_{i_k})$$

- Aim: inform user about interesting new information
e.g. personalized news service
- Method:
 1. User creates initial interest profile (= query)
 2. Each new document is compared to profile
 3. If similarity is "high enough", select and forward document to user
 4. Refine query using user feedback
 5. New profile = $\alpha \times$ old profile +
 $\frac{\beta}{\#rel.docs.} \times \sum$ relevant docs -
 $\frac{\gamma}{\#non-rel.docs.} \times \sum$ non-relevant docs
 6. Intuitively, add terms occurring in many relevant documents, remove terms occurring in many non-relevant documents

Motivation:

- Manual thesauri are:
 - general purpose (Roget's Thesaurus, WordNet) – difficult to use for document retrieval
 - retrieval-oriented (INSPEC, MeSH) – expensive to build and maintain
- Construct an automatic thesaurus (based on information about co-occurrence of words in a collection)

Text Mining: Word Relations

- **Association:** if two terms co-occur within the same paragraph, they constitute an association

$\langle \text{term}_1, \text{term}_2, \text{assoc. frequency} \rangle$

- Gather data about term-associations over a large amount of text
- Refine associations:
 - Discard associations with frequency 1
 - Discard terms that are associated with too many other terms (people, state, company, etc.)

Text Mining: Word Relations

- Each term is represented by a vector of associated terms

$$T = (\langle t_1, w_1 \rangle, \dots, \langle t_n, w_n \rangle)$$

⇒ term = pseudo document

- Compare query to the term vectors (instead of document vectors)

$$Sim(Q, T) = \sum_i wt(q_i) \times wt(t_i)$$

- Most “similar” terms are added to the query
- Example: 1986 US Immigration Law
 - similar terms: *illegal immigration, amnesty program, simpson-mazzoli*

Text Mining: Word Relations

Experimental results:

- Data: 500,000 documents (news, computer abstracts, govt. documents); 50 queries
- Baseline average precision: 37%
- Improves to 6 - 30% by using thesaurus
- 2 weeks to generate association data!
- Processing time can be reduced without major loss in performance by using a subset of the document collection

Text Mining: Classification

- Users may prefer browsing through document collection instead of doing a direct keyword search
- Search sites organize web-pages into hierarchy of subject categories
e.g. *Science > Physics > Relativity*
- New web-pages need to be inserted into appropriate class automatically

Text Mining: Classification

- Training: initially, documents are classified manually
- Class vector computed for each class based on the documents contained in that class
e.g. $D_{CS_1} \Rightarrow$ algorithm, complexity, graph
 $D_{CS_2} \Rightarrow$ searching, text, algorithm
 $D_{CS} \Rightarrow$ algorithm(2), complexity, ...
- New document compared to class vectors at each level of hierarchy to determine best fit
Example:
 1. compute
 $Sim(D, CS), Sim(D, Physics), Sim(D, Maths), \dots$
 2. compute
 $Sim(D, relativity), Sim(D, optics), Sim(D, mechanics), \dots$

Text Mining: Document Relationships

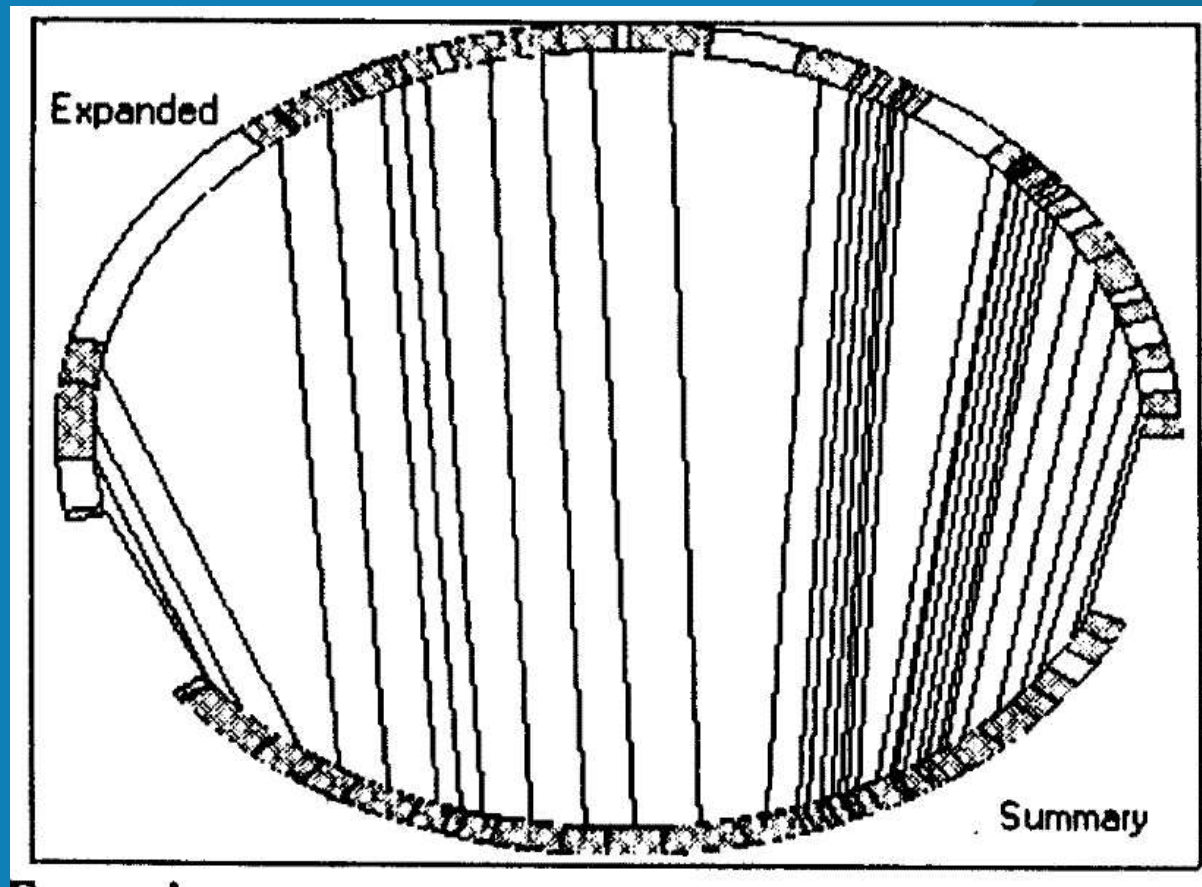
- Creator of a web-page may not provide links to other important related pages
- Links between related pages should be automatically discovered
 - Related pages are expected to have a high similarity
- Type of relationship should be detected if possible
 - *summary-expansion, generalization-specialization, etc.*

Text Mining: Document Relationships

- Break each document into paras
- Construct document relationship graph
 - nodes - paras
 - edges - join paras that have a high similarity
- Depending on patterns in the graph, likely relationship may be detected

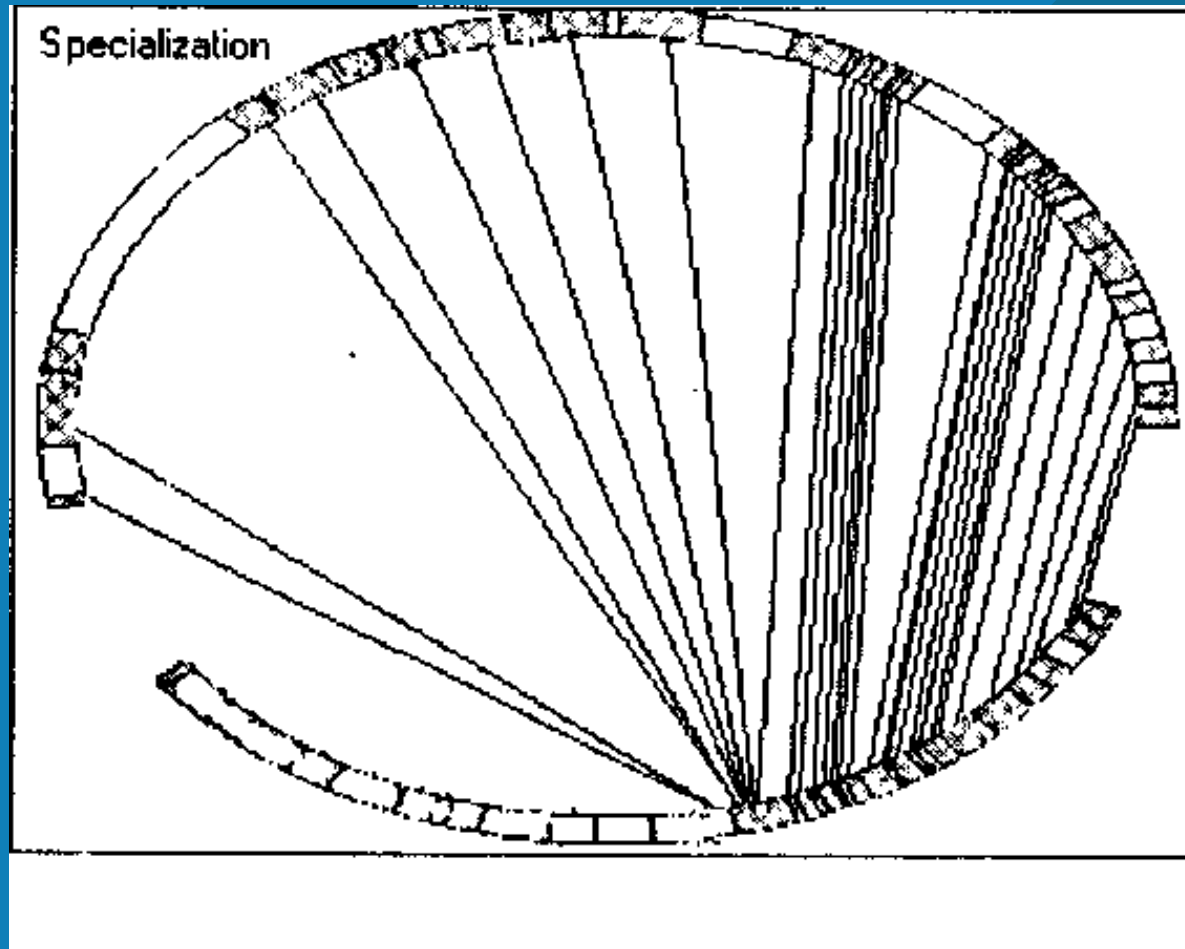
Document Relationships

- Summary-Expansion
 - *news-in-brief vs. news-in-detail*



Document Relationships

- Generalization-Specialization
 - *unmanned space missions vs. Pioneer 10*



Text Mining: Summarization

- Manual summarization method:
 - read text and understand it
 - extract salient points
 - write the summary
- Automatic approximation:
 - Break document into paragraphs
 - Compute para-para similarities
 - Construct document relationship graph
 - Extract “important” paras (expected to have high degree)
 - *Bushy* paras: paras connected to many other paras
 - *Depth first* paras: from starting para, go to most similar para
- Comprehensiveness vs. coherence
 - comprehensive: covers salient points
 - coherent: easy to read

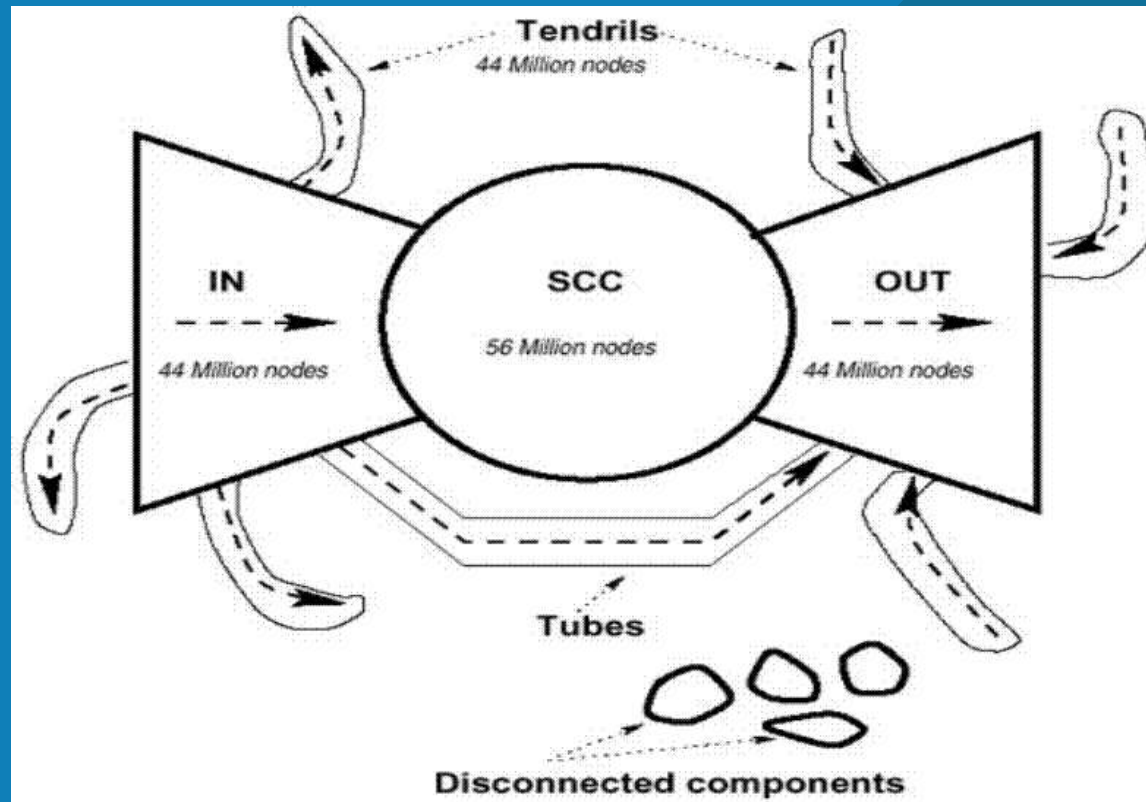
Evaluation:

- Manually extract best paras; compute overlap between automatic and manual extract
- Problems:
 - agreement between humans is low (60%)
 - just choosing first few paras works well

Structure Mining: Outline

- Web as a graph
- Detecting hubs and authorities
- Page ranking (Google)
- Community detection

- Web is a **directed graph**: set of pages (nodes) connected by hyperlinks (edges)

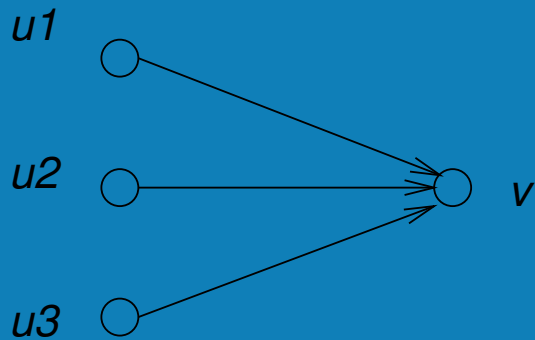


- Based on about 200 million pages, 1.5 billion links
- One strongly connected component (path from each node to every other node)
- IN – set of newly formed nodes with outgoing links into the centre
- OUT – introvert nodes with only incoming links (e.g. corporate and e-commerce sites) from the centre
- Tendrils and tubes (nodes in IN-tendrils connect to nodes in OUT-tendrils)
- Randomly chosen pair is connected only 24% of the time (average distance 16)

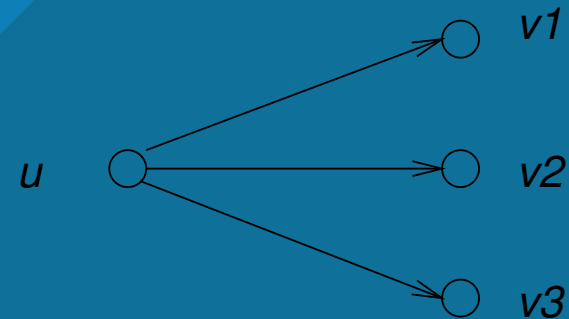
- **Hyperlink Induced Topic Search**
 - Kleinberg, 1998
- Identification of
 - **authorities** – authoritative, high-quality web pages on broad topics
 - **hubs** – web pages that link to a collection of authorities
- A good authority is pointed to by many good hubs
- A good hub points to many good authorities
- Inspired by the study of social networks and citation analysis

Structure Mining: HITS

- **Root set:** given a broad query, collect the N highest ranked pages for the query from a text-based search engine
- **Expanded set:** add pages pointing to pages in root set, and pages pointed to by pages in root set
- Iteratively update authority and hub scores



$$a(v) = h(u1) + h(u2) + h(u3)$$



$$h(u) = a(v1) + a(v2) + a(v3)$$

HITS: Problems, Solutions

■ Problems:

- Clique attacks (www.411fun.com, 411fashion.com,) etc.
- Mixed hubs and topic drifting

■ Solutions:

- make use of *anchor text* (the text surrounding a link) and boost weight of links which occur near instances of query terms
- eliminate outliers from the expanded set
- partition mixed hubs into segments

Structure Mining: PageRank

- Used in Google Search Engine
- 'Global' ranking of every web page calculated based on hyperlink structure of web (content ignored)
- Documents with matching keywords are returned in the global rank order
- Principle: Highly linked pages are more important than pages with a few links
A page has a high rank if the sum of the ranks of its back-links is high
- Most effective for underspecified (general) queries

Structure Mining: Web Communities

- **Community:** group of web pages sharing a common interest
 - Explicit: Yahoo, Google, etc.
 - Implicit: have to be discovered using content + hyperlinks
- Similarity
 - Method 1: A and B are related if one links to the other
 - Method 2: A and B are related if
 - a number of pages contain links to both A and B
 - A and B both link to a number of pages
- Use matrix algebra methods (PCA, eigenvalue analysis), or graph theoretic methods (community trawling)

Usage Mining: Outline

- Premises and goals
- Criteria for success
- Architecture
- Opinion mining
- Amazon

Usage Mining: Data Sources

- Server logs
 - For each access, web servers register an entry in a log file containing requesting host, user id, timestamp, page requested, browser type, referring page, etc.
- Packet sniffer logs
 - monitor network traffic and extract usage data directly from TCP/IP packets
- User sessions/queries
- User profiles, registration data, bookmarks

Usage Mining: Applications

- Enhance server performance (caching, prefetching)
- Improve web site navigation (general / customized)
- Identify potential customers for e-commerce
- Advertising:
 - identify potential prime advertisement locations
 - targeted advertising

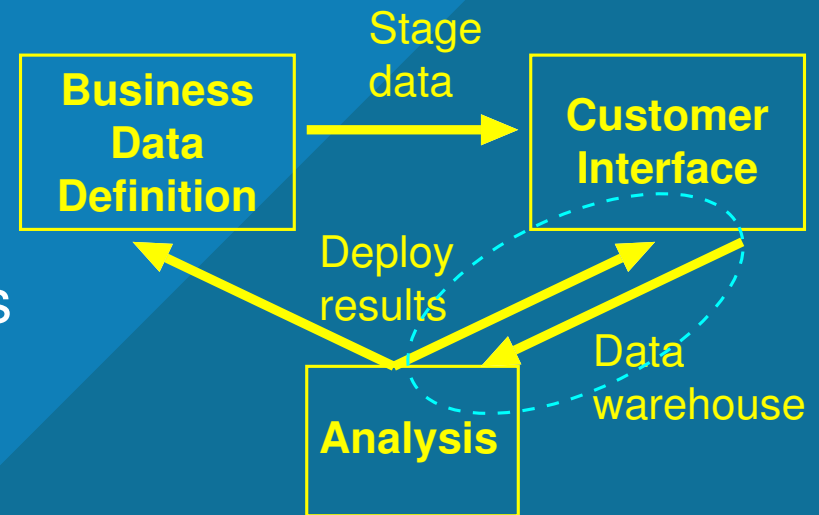
Usage Mining: Desiderata

- Rich data
 - wide customer records with many potentially useful fields allow data mining algorithms to search beyond obvious correlations
 - recording the actions of customers in the virtual store is much easier (items examined, selected, purchased)
- Large volumes of data
 - required to train reliable (complex) models
- Controlled/reliable data collection
 - manual data entry / integration from legacy systems avoidable
- Evaluation of return on investment
- Ease of integration

Usage Mining: Integrated Architecture

■ Customer Interface

- data collector needs to be integrated into interface
- sale transactions + other details (redirection, promotion, personalization, etc.)



■ Business Data Definition

- merchandise-related information (products, price, etc.)
- content information (web page templates, articles, images, multimedia)
- business rules (promotions / personalization / cross-selling rules)

!! important to have rich set of metadata attributes

Usage Mining: Data Collection

- Server/packet sniffer logs: non-intrusive but very low-level
- Problems:
 - User identification
 - user may use different machines/browsers
 - use of public access PCs / proxy servers / caching
 - ⇒ use login ids, cookies, “negative” expiration dates, etc.
 - Session identification: HTTP is “stateless”
 - login ids – painful to register at sites
 - ⇒ use timeouts (30 minutes)
 - Intra-page navigation
 - Data generated by CGI scripts, dynamic pages
 - Encryption / secure pages (for packet sniffers)

Usage Mining: Data Collection

- Application server logs: can collect high-level information
 - application server has detailed knowledge of content sent to user
 - server can use cookies or URL encoding to keep track of sessions, events, user identities
- High-level information can be used to calculate:
 - micro-conversion rates: for each step of the purchasing process, the fraction of products that are successfully carried through to the next step of the purchasing process
view → add to cart → checkout
 - effectiveness of personalization: correlation between using a personalization rule and shopping cart/checkout events compared to using control groups

- Aggregation: needed to convert collected data into forms that are more amenable to analysis
 - examples:
 - how much money does a customer spend on books?
 - what is the frequency of a customer's purchases?
 - what kind of shipping options are chosen what portion of the time?
 - may not be easy to achieve using standard aggregation tools provided by SQL, etc.
- Transformation of dates:
 - difference between order date and ship date
 - extract day of the week, month, quarter, season, etc.

- Basic reporting:
 - what are the top/worst selling products?
 - what are the top successful / failed searches?
 - who are the top referrers by visit count / sales amount? (*)
 - what are the top abandoned products? (*)
 - what is the distribution of web browsers?
- Visualization tools
- Association rule mining
- Classifiers (Bayesian, decision tree, etc.)
- Interactive model interpretation / modification tools are a must

Feature-based opinion summarization

- Identify the features of the product that customers have expressed opinions on (called opinion features)
- For each feature, identify how many customer reviews are positive / negative

Examples:

The pictures are very clear.

Overall a fantastic, very compact, camera.

While light, it will not easily fit in pockets. (HARD!)

Feature identification

1. POS tagging + chunking: identify nouns, verbs, adjectives, simple noun groups, verb groups
2. Transaction creation for each sentence: item \equiv normalized nouns / noun phrases
3. Association rule mining: all itemsets with $> 1\%$ support are candidate *frequent features*
4. Feature pruning:
 - keep features that have some *compact* occurrences
 - keep singleton itemsets only if they occur enough times in isolation
e.g. *manual* vs. *manual mode*, *manual setting*
5. Infrequent feature identification: noun/noun phrase that occurs closest to a known *opinion word*

Sentiment / orientation identification

1. Examine each sentence in the review database
2. If it contains a frequent feature, extract all the adjective words as opinion words
3. For each feature in the sentence, the nearby adjective is recorded as its *effective opinion*
4. Look up adjective in a list of adjectives with known orientation, or consult WordNet (discard unknowns)
 - adjectives arranged in bipolar structures

file://localhost/mnt/pen-ext2/narnia/narnia3.html 100%

amazon.com **Mandar's Store** Books See All 34 Product Categories Your Account | Cart | Your Lists | Help |

Advanced Search | Browse by Subject | Bestsellers | The New York Times® Best Sellers | Magazines | Corporate Accounts | Amazon Shorts | Amazon Connect | Bargain Books | Textbooks

Search Books GO Find Gifts Web Search GO

Join **Amazon Prime** and ship Two-Day for free and Overnight for \$3.99.

SEARCH INSIDE!™



The Chronicles of Narnia (Narnia) (Paperback)
 by [C. S. Lewis](#), [Pauline Baynes](#) (Illustrator) "This is a story about something that happened long ago when your grandfather was a child..."
[\(more\)](#)
Explore: [Citations](#) | [Books on Related Topics](#) | [CAPs](#)
Browse: [Front Cover](#) | [Copyright](#) | [Table of Contents](#) | [Excerpt](#) | [Back Cover](#) | [Surprise Me!](#)

Add to Shopping Cart
 or
Buy now with 1-Click®

Ship to:
 Sunnyvale- CA
 Add gift-wrap/note

A9.com users save 1.57% on Amazon.
[Learn how.](#)

[Share your own customer images](#)
[Search inside this book](#)

List Price: \$19.99
Price: **\$12.99** & eligible for **FREE Super Saver Shipping** on orders over \$25. [Details](#)
You Save: \$7.00 (35%)

Availability: Usually ships within 24 hours. Ships from and sold by Amazon.com.

Want it delivered Friday, July 21?
 Choose **One-Day Shipping** at checkout. [See details](#)

141 used & new available from \$3.00

More Buying Choices
141 used & new from \$3.00

Available for in-store pickup now from: \$19.99
 Price may vary based on availability
 Enter your ZIP Code: GO

Have one to sell?
[Sell yours here](#)

[Add to Wish List](#)
[Add to Shopping List](#)

Better Together

Buy this book with [A Family Guide to Narnia: Biblical Truths in C.S. Lewis's the Chronicles of Narnia](#) by Christin Ditchfield today!



+



Buy Together Today: \$23.38



Buy both now!

Customers who bought this item also bought

[A Family Guide to Narnia: Biblical Truths in C.S. Lewis's the Chronicles of Narnia](#) by [Christin Ditchfield](#)

[The Lion, the Witch and the Wardrobe \(Full-Color Collector's Edition\)](#) by [C. S. Lewis](#)

[A Wrinkle in Time](#) by [Madeleine L'Engle](#)

[Harry Potter Paperback Boxed Set \(Books 1-5\)](#) by [J. K. Rowling](#)

[Companion to Narnia, Revised Edition: A Complete Guide to the Magical World of C.S. Lewis's The Chronicles of Narnia](#) by [Paul F. Ford](#)

Explore similar items: in [Books](#), in [DVD](#)

What do customers ultimately buy after viewing items like this?

75% buy [The Quillan Games \(Pendragon\)](#) by [D. J. MacHale](#) ★★★★★ \$10.85

20% buy the item featured on this page: [The Chronicles of Narnia \(Narnia\)](#) by [C. S. Lewis](#) ★★★★★ \$12.99

5% buy [The Chronicles of Narnia Movie Tie-in Box Set \(rack\) \(Narnia\)](#) by [C. S. Lewis](#) ★★★★★ \$29.70

[Compare these items](#)

[Explore Similar Items](#)

Your Recent History

[Learn more](#)

Recently Viewed Products



[The Chronicles of Narnia Boxed Set](#) by C.S. Lewis

Recent Searches

- ["chronicles of narnia"](#)
- ["chronicles of narnia"](#)

Customers who bought items in your Recent History also bought:



[A Family Guide to Narnia](#) by Christin Ditchfield



[The Lion, the Witch and the Wardrobe](#) by C. S. Lewis

[Visit the Page You Made](#)

Editorial Reviews

Amazon.com

The Chronicles of Narnia, by C.S. Lewis, is one of the very few sets of books that should be read three times: in childhood, early adulthood, and late in life. In brief, four children travel repeatedly to a world in which they are far more than mere children and everything is far more than it seems. Richly told, populated with fascinating characters, perfectly realized in detail of world and pacing of plot, and profoundly allegorical, the story is infused throughout with the timeless issues of good and evil, faith and hope. This boxed set edition includes all seven volumes. --*This text refers to an out of print or unavailable edition of this title.*

Book Description

Beloved by generations for more than 50 years, this classic children's series is now available in a special adult edition.

[See all Editorial Reviews](#)

Spotlight Reviews

[Write an online review](#) and share your thoughts with other customers.

[Search Customer Reviews](#) ([What's this?](#))

GO!

2624 of 2881 people found the following review helpful:

★★★★☆ **Don't Tamper With Perfection**, December 9, 2002

Reviewer: **C. N. White "neilicus107"** (Raleigh, NC) - [See all my reviews](#)

REAL NAME™

(please note that this review concerns only the new publications)

The Chronicles of Narnia are perfect books. They are wonderful for children and adults, and can be

Customer Reviews

Average Customer Review: ★★★★★

[Write an online review](#) and share your thoughts with other customers.

0 of 1 people found the following review helpful:

★★★★★ **A Gift from Narnia**, July 18, 2006

Reviewer: **KitCat** - [See all my reviews](#)

Purchasing the book set was the best thing I ever did. The color illustration on each book in the series sets an image in the reader's mind of what to expect, that is sets the stage for the story to come. The illustrations are very well done. My Mom started reading the books and enjoyed them so much, I bought her a set. The truth of the matter is, it was not only for her enjoyment, but also, because the books are something a person doesn't want to loan out to just anyone. On top of everything else, they're a bargain. They're truly a gift from Narnia.

Was this review helpful to you? ([Report this](#))

Customer Discussions Beta ([what's this?](#))

[?](#) [Help](#)

Amazon customers talk about this product and related topics.

Related forums

12 discussions in 2 forums

This product's forum

(11)

[Chronicles of Narnia \(1\)](#)

Listmania!

Search Listmania!

GO!



[Books I Own](#): A list by [avidreadertoo "avt"](#)



[List Of Books I Must Read =\)](#): A list by [Wind Dancer](#)



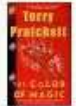
[My Classic Reading List](#): A list by [Hattie Forrester](#)

[Create a Listmania! list](#)

So You'd Like to...



[read the children's books that are good for adults](#): A guide by [Joseph Dewey](#), a good reader



[Fantasy books being made into movies](#): A guide by [H. Maher](#), super duper reader



[become a humanist Christian, 3 of 4](#): A guide by [Terry Bohannon](#), an intellectually curious English major.

[Create your guide](#)

Look for similar items by category

[Book Clubs](#) > [Literature & Fiction](#) > [Classics](#)

[Subjects](#) > [Children's Books](#) > [Ages 9-12](#) > [General](#)

[Subjects](#) > [Children's Books](#) > [Authors & Illustrators, A-Z](#) > (L) > [Lewis, C.S.](#)

[Subjects](#) > [Children's Books](#) > [Literature](#) > [Classics by Age](#) > [General](#)

[Subjects](#) > [Children's Books](#) > [Literature](#) > [Science Fiction, Fantasy, Mystery & Horror](#) > [Science Fiction, Fantasy, & Magic](#)

[Subjects](#) > [Children's Books](#) > [Religions](#) > [Christianity](#) > [Ages 9-12](#)

[Subjects](#) > [Children's Books](#) > [Religions](#) > [Christianity](#) > [Series](#) > [The Chronicles of Narnia](#)

[Subjects](#) > [Literature & Fiction](#) > [General](#) > [Classics](#)

Look for similar items by subject

- [Children](#)
- [Children's 9-12 - Literature - Classics / Contemporary](#)
- [Children's Books/Ages 9-12 Fiction](#)
- [Children: Grades 4-6](#)
- [Classics](#)
- [Fantasy](#)
- [Good and evil](#)
- [Juvenile Fiction](#)
- [Science Fiction, Fantasy, & Magic](#)
- [Christian Interest](#)
- [Fiction / Classics](#)
- [Juvenile Fiction / Religious / General](#)
- [Reading Group Guide](#)

Find books matching ALL checked subjects

i.e., each book must be in subject 1 AND subject 2 AND ...

- Web Mining Research: A Survey, R. Kosala, H. Blockeel, *SIGKDD Explorations*, 2(1), July, 2000.
- Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, J. Srivastava, R. Cooley, M. Deshpande, P. Tan, *SIGKDD Explorations*, 1(2), Jan, 2000.
- WEBKDD 2000 Worskhop on Web Mining for E-Commerce – Challenges and Opportunities.
<http://robotics.stanford.edu/~ronnyk/WEBKDD2000/papers/>
- Data Mining and Knowledge Discovery, 5 (2001), 6 (2002).
- Communications of the ACM, 45(8), August 2002.
- Machine Learning, 57, 2004.
- 14th International World Wide Web Conference (WWW 2005) Tutorial on Web Content Mining (Bing Liu)
<http://www.cs.uic.edu/~liub>