Web Mining


Web
Mining
Analysis
Text
Data
Information
Applications
Social
Machine
based
network
discovery
Detection
mining
Link
Security
Address
Boilerplate
Influential
Features
Terrorism
Discovery
Variable
patterns
issues
systems
Systems
Providers
CoAd
Dynamics
DBpedia
Optimized
Source
recommender
approach
social
Spam
Multiple
Regularization
Helps
Hub
Scale
App
Query
Large-Scale
Conflicting
Views
Reading
Networks
Temporal
Supporting
Counter
Targeting
business
model
Diverse
analysis
funnel
Latent
Shallow
Learning
Document
Enterprises
Identifying
Bloggers
Android
Classification
State
Environmental
technologies
graphical
Unified
Computational
Linked
Related
sites
Recommender
advertising
Truth
Overload
Semantic
Current
indexes
Behavioral
models
communication
Sensing
Understanding
Heterogeneous
Using
View
Art
pajek
users
Articles
Models

Contents

References

- https://pdfs.semanticscholar.org/a408/c8224937319b9a45446f33696d6b7d74140f.pdf
- https://en.wikipedia.org/wiki/Web_mining
- https://www.educba.com/data-mining-vs-web-mining/
- http://www.ijeert.org/pdf/v3-i10/9.pdf
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8303127
- http://dmr.cs.umn.edu/Papers/P2004_4.pdf
- https://paginas.fe.up.pt/~ec/files_0910/slides/aula_6_WebMining.pdf

References (cont.)

- https://www.slideshare.net/AmirFahmideh/web-mining-structure-mining
- https://www.slideshare.net/ErJagratGupta/web-mining-presentation-final?qid=56467646-d8a4-45e3-8c16-2b32c3275f4a&v=&b=&from_search=9
- https://www.slideshare.net/MuditDholakia/web-mining-47715019?next_slideshow=1
- https://www.cs.uic.edu/~liub/WebContentMining.html
- https://paginas.fe.up.pt/~ec/files_0910/slides/aula_6_WebMining.pdf

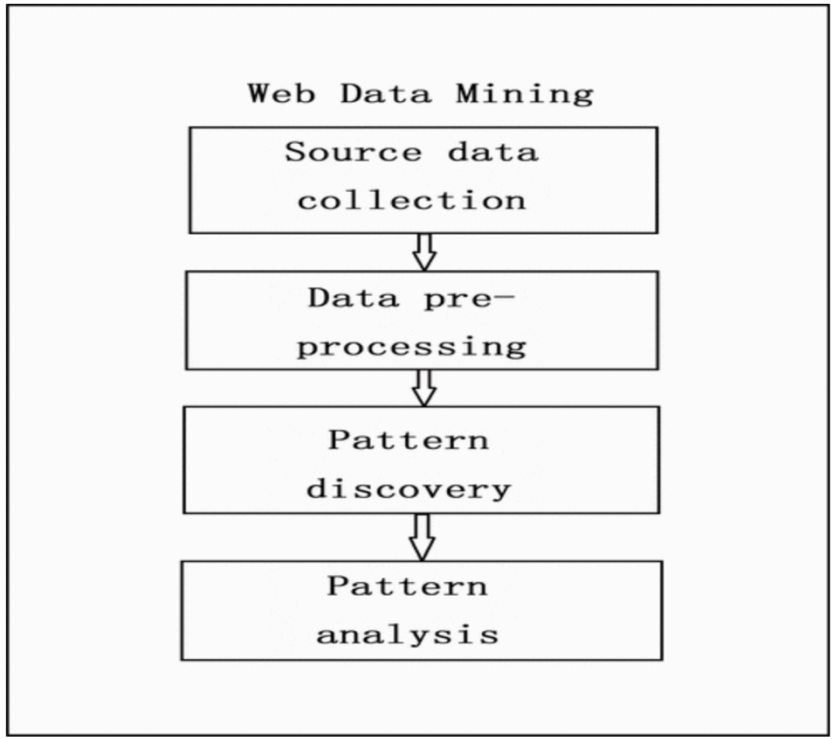
What is Web Mining?

- Web mining is the use of data mining techniques to extract knowledge from web data.
- Web data includes :
  - web documents
  - hyperlinks between documents
  - usage logs of web sites
- The WWW is huge, widely distributed, global information service centre and, therefore, constitutes a rich source for data mining.

Data Mining vs Web Mining

- **Data Mining :** It is a concept of identifying a significant pattern from the data that gives a better outcome.
- **Web Mining :** It is the process of performing data mining in the web. Extracting the web documents and discovering the patterns from it.

Web Data Mining Process

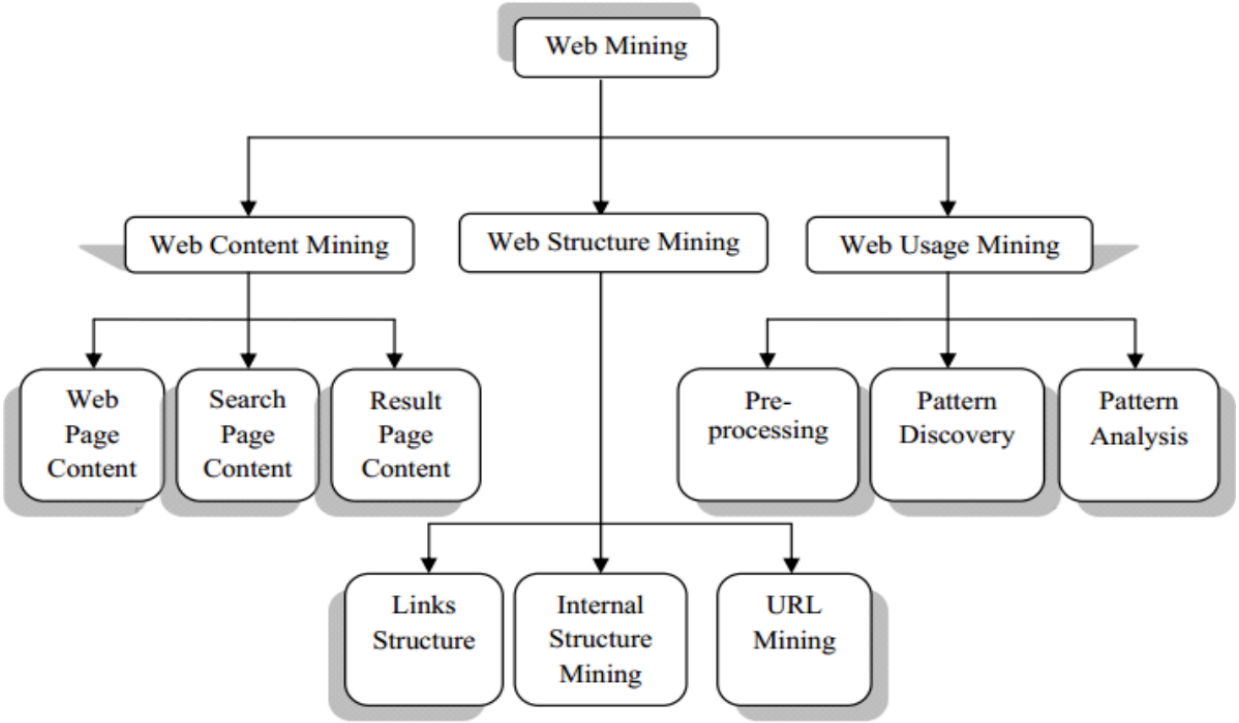


https://ieeexplore.ieee.org/document/5485404/

Issues

- Web data sets can be very large
  - Tens to hundreds of terabyte
- Cannot mine on a single server
  - Need large farms of servers
- Proper organization of hardware and software to mine multi-terabyte data sets
- Difficulty in finding relevant information
- Extracting new knowledge from the web

Web Mining Taxonomy



https://www.researchgate.net/figure/Taxonomy-of-Web-Mining-Source-Kavita-et-al-2011_fig1_282357293

Web Content Mining - Introduction ??

- Mining, extraction and integration of useful data, information and knowledge from Web page content.
- Web content mining is related but different from data mining and text mining.
- Web data are mainly **semi-structured** and/or **unstructured**, while data mining deals primarily with structured data.

Web Content Mining Includes ? ? ?

Image ref-1
image-ref-2
image-ref-3

Unstructured Web Data Mining

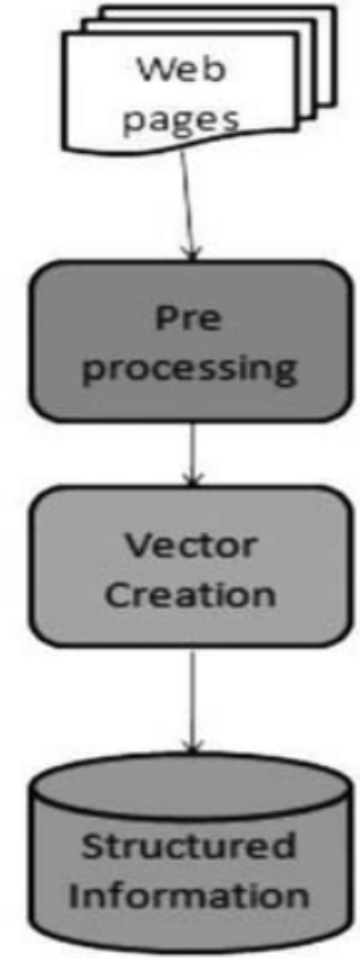


Image Source http://www.3idatascraping.com/6-tips-on-how-to-do-data-scraping-of-unstructured-data.php
https://www.researchgate.net/figure/The-Progress-of-Web-Content-Mining_fig13_299529651

Unstructured Documents - Feature Extraction

- Bag of words to represent unstructured documents
  - Takes single word as feature
  - Ignores the sequence in which words occur
- Features could be
  - Boolean
    - Word either occurs or does not occur in a document
  - Frequency based
    - Frequency of the word in a document
- Variations of the feature selection include
  - Removing the case, punctuation, infrequent words and stop words etc..
- Features can be reduced using different feature selection techniques:
  - Information gain, mutual information, cross entropy.
  - Stemming: which reduces words to their morphological roots.

Structured Web Data

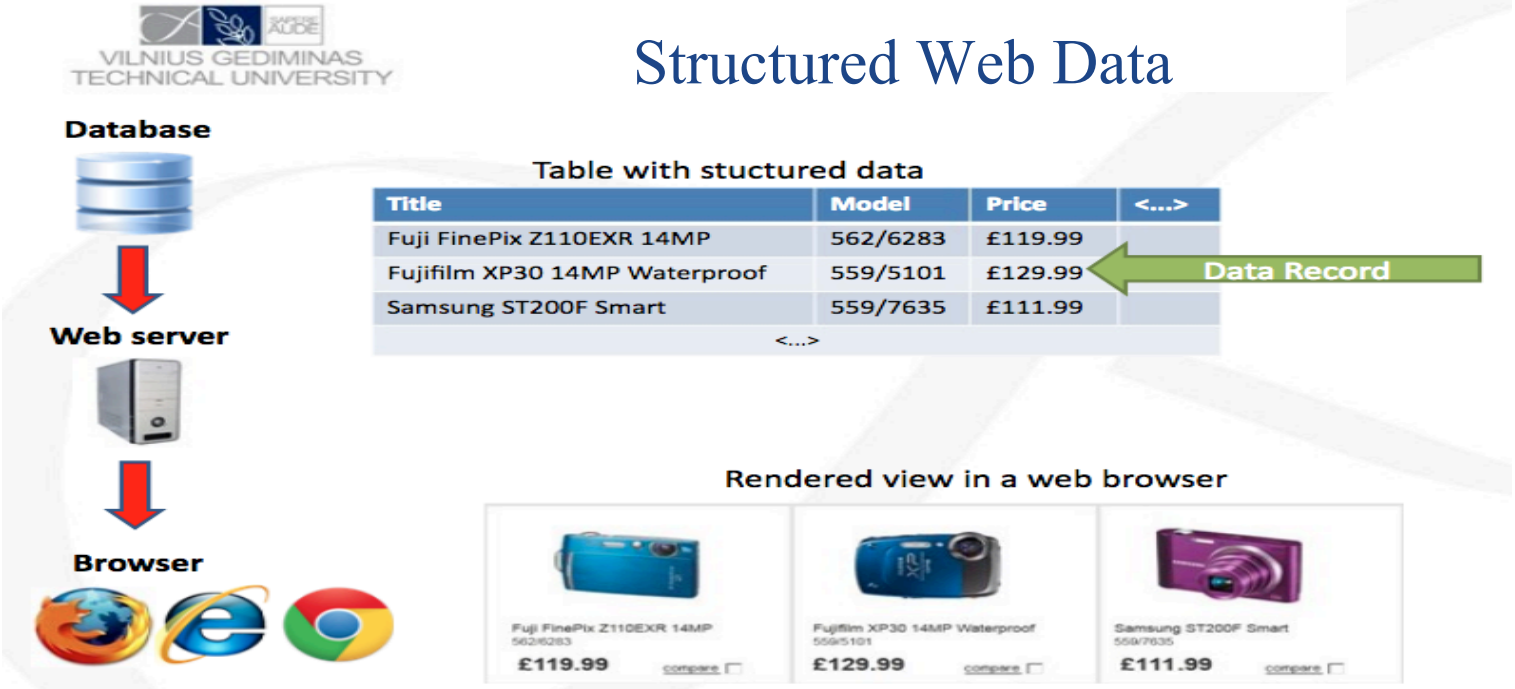


Image Source: http://slideplayer.com/slide/6356153/

Mining Techniques Using Agent and Database

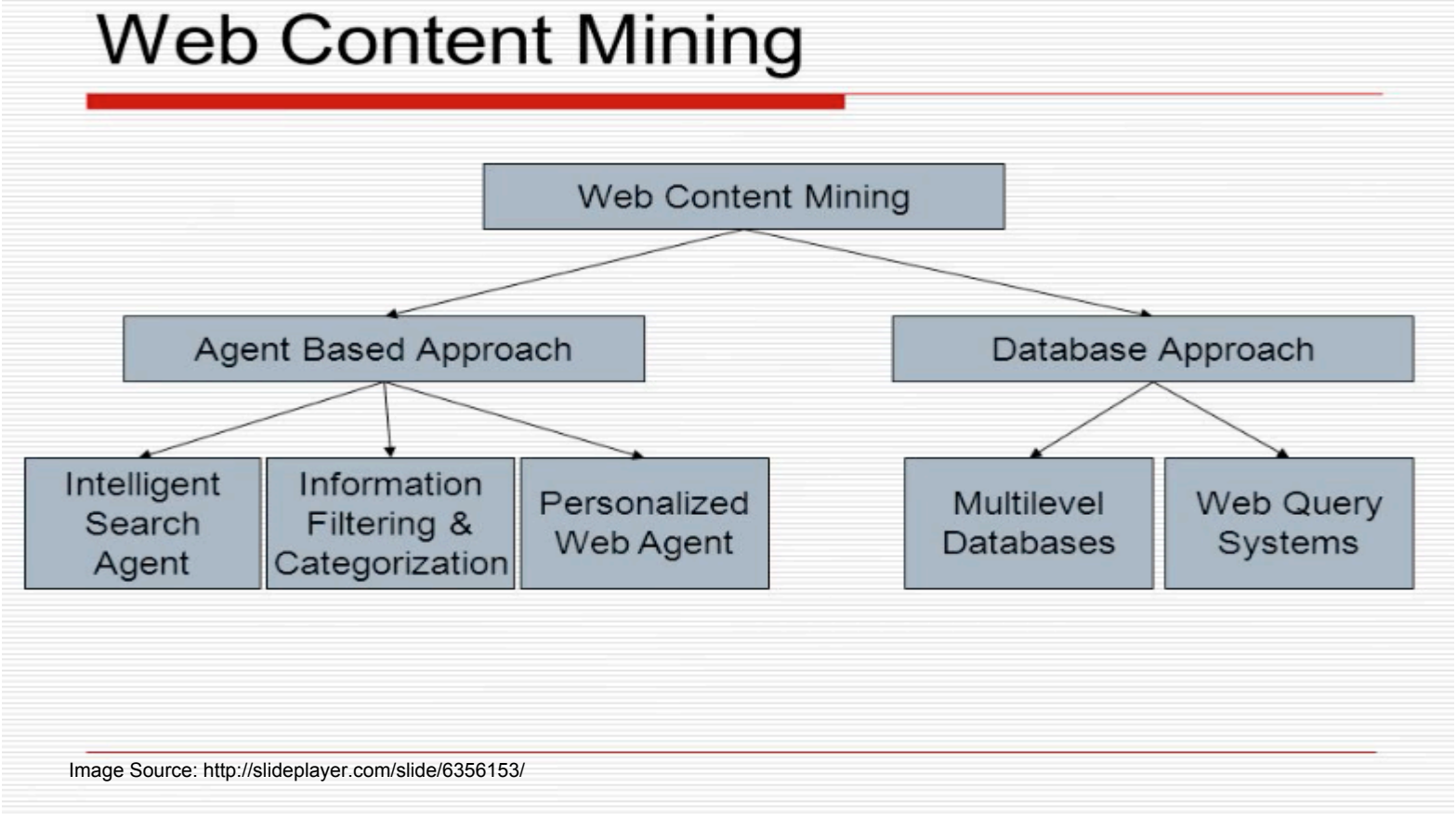


Image Source: http://slideplayer.com/slide/6356153/

Agent-Based Approach

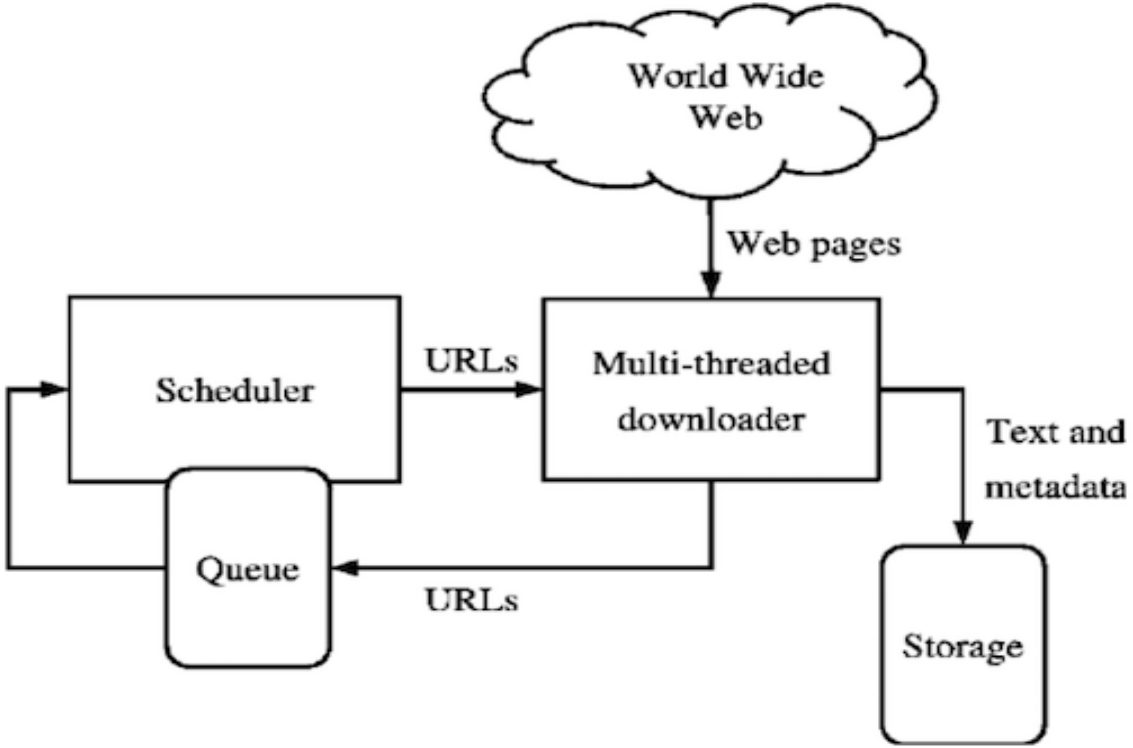
- **Intelligent-Search-Agents** developed that searches for characteristics to organize and interpret the discovered information.
- **Information-Filtering/Categorization** - Using various information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them. HyPursuit, BO (Bookmark Organizer).
- Development of **sophisticated AI systems** acting on behalf of users autonomously or semi-autonomously to discover and organize information.

Database Approaches

Used for **transforming unstructured** data **into** more **structured** and h**igh-level collections** of resources, such as in **relational databases**, and using standard database **querying** mechanisms and data **mining** techniques to access and **analyze** this information.

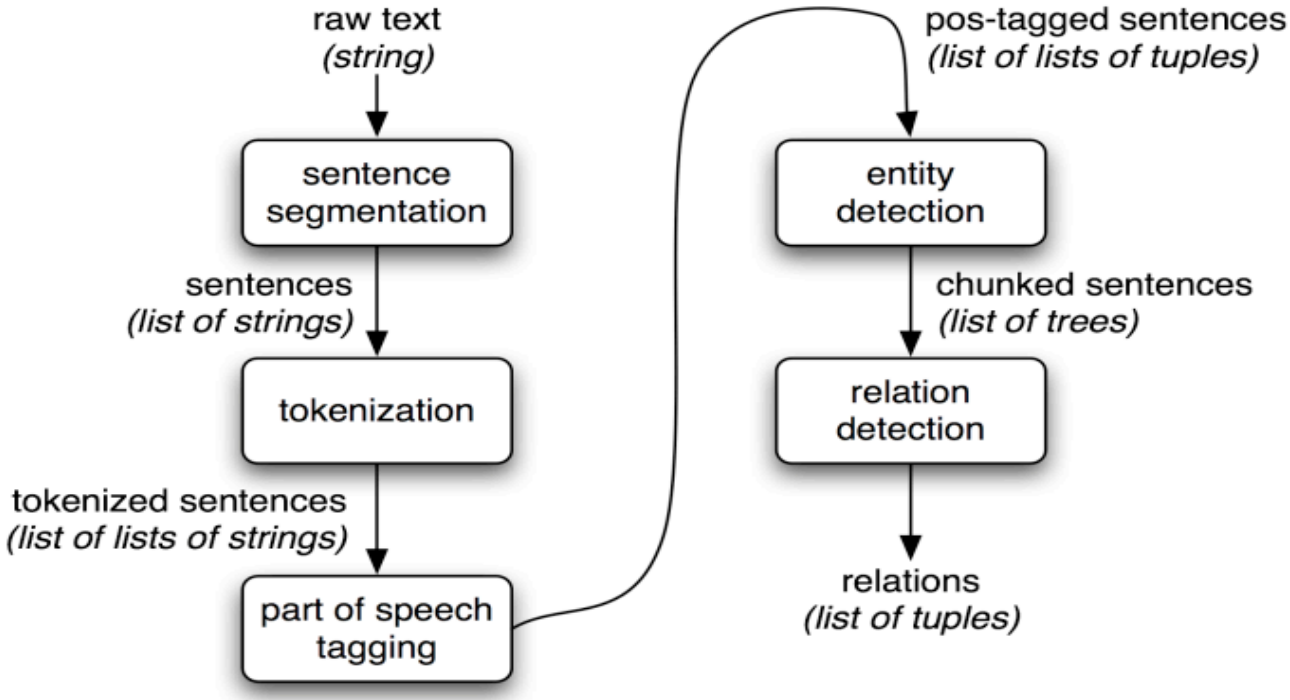
- **Multilevel**-Databases
  - **lowest** level - semi- structured information is kept
  - **High** level - generalizations from lower levels organized into relations and objects.
- **Web-Query** Systems
  - Web-based query systems and languages developed such as SQL, NLP for extracting data.

Typical Crawler



Img Source: https://www.researchgate.net/figure/Typical-high-level-architecture-of-a-Web-crawler_fig4_228548882

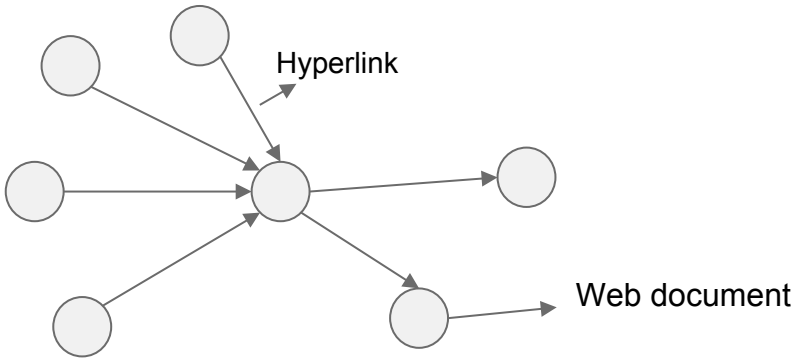
Text Mining - Brief



Img source: https://edumine.wordpress.com/2015/09/14/part-of-speech-tags-in-text-mining/

What is Web Structure Mining?

- Web structure mining is the process of discovering structure information from the web.
- The structure of typical web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages.



Web Structure Mining (cont.)

- This type of mining can be performed either at the document level(intra-page) or at the hyperlink level(inter-page).
- The research at the hyperlink level is called Hyperlink analysis.
- Hyperlink structure can be used to retrieve useful information on the web.

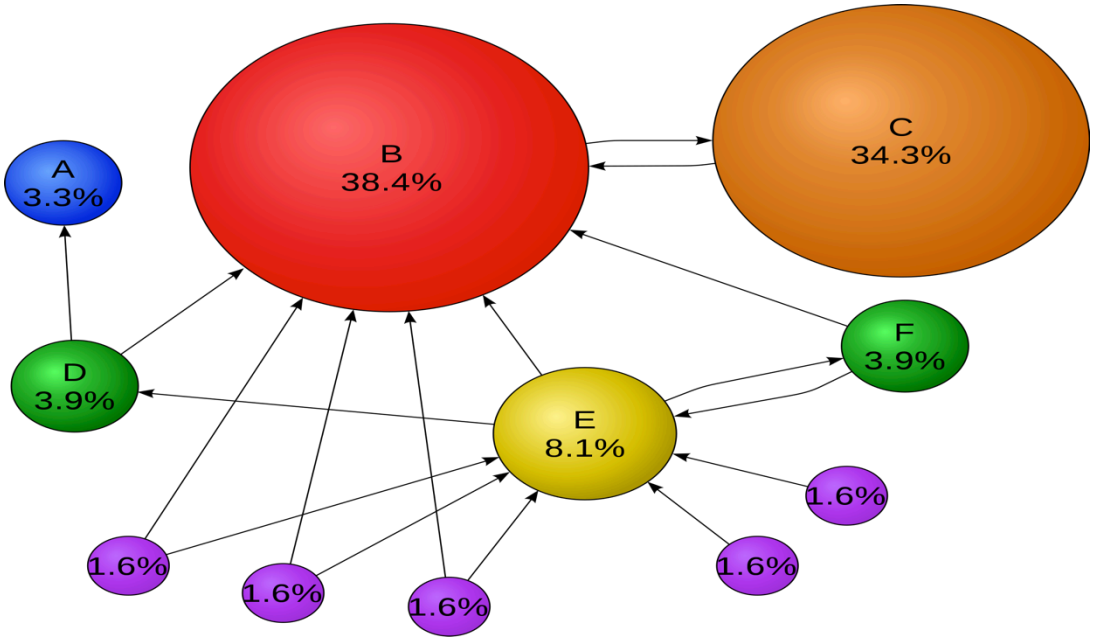
There are two main approaches:

- PageRank
- Hubs and Authorities - HITS

PageRank

- Used to discover the most important pages on the web.
- Prioritize pages returned from search by looking at web structure.
- Importance of pages is calculated based on the number of pages which point to it (backlinks).
- Weighting is used to provide more importance to backlinks coming from important pages.
- PR(p) = (1-d) + d (PR(1)/N1 + …… + PR(n)/Nn)
  - PR(i): PageRank for a page i which points to target page p.
  - Ni: Number of links coming out of page i.
  - d: constant value between 0 and 1 used for normalization.
  - (1-d): Bit of probability math magic so that sum of all webpages pageranks should be one.

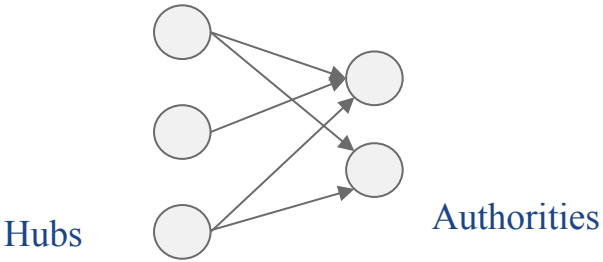
PageRank (cont.)



https://en.wikipedia.org/wiki/PageRank

Hubs and Authorities

- Authoritative pages
  - Authors defines an authority as the best source for the request.
  - Highly important pages.
  - Best source for requested information.
- Hub pages
  - Contains links to highly important pages.



HITS (Hyperlink Induced Topic Search)

- Iterative algorithm for mining the Web graph to identify the topic hubs and authorities.
- Algorithm:
  - Let’s consider a matrix A with rows and columns corresponding to web pages. Aij =1 indicates that page i links to j and 0 otherwise.
  - Let a and h are vectors, whose ith component corresponds to the degree of authority and hubbiness of ith page.
  - Hubbiness of the page is defined as the sum of the authorities of all the pages it links to. i.e h = A x a.
  - Authority of the page is defined as the sum of hubbiness of all the pages that link to it. i.e. a = At x h. where At is the transposed matrix.

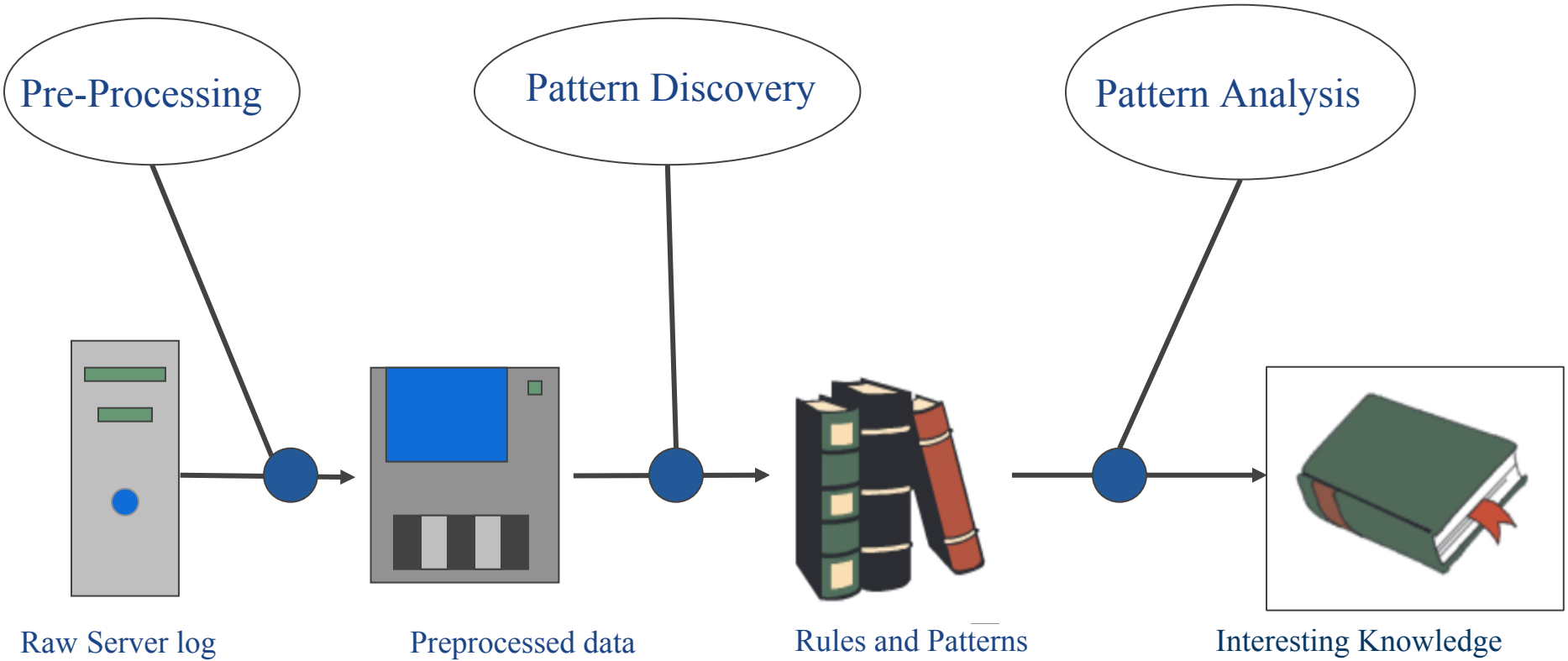
Web Structure Mining applications

- Information retrieval in social networks.
- To find out the relevance of each web page.
- Measuring the completeness of Web sites.
- Used in search engines to find out the relevant information.

Web Usage Mining

- **Web usage mining**: automatic discovery of patterns in clickstreams and associated data collected or generated as a result of user interactions with one or more Web sites.
- **Goal**: analyze the behavioral patterns and profiles of users interacting with a Web site.
- The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common interests.
- Data in Web Usage Mining:
  a. Web server logs
  b. Site contents
  c. Data about the visitors, gathered from external channels

Three Phases



reference

Data Preparation

- **Data cleaning**
  - By checking the suffix of the URL name, for example, all log entries with filename suffixes such as, \gif, jpeg, etc
- **User identification**
  - If a page is requested that is not directly linked to the previous pages, multiple users are assumed to exist on the same machine
  - Other heuristics involve using a combination of IP address, machine name, browser agent, and temporal information to identify users
- **Transaction identification**
  - All of the page references made by a user during a single visit to a site
  - Size of a transaction can range from a single page reference to all of the page references

Pattern Discovery Tasks

- **Clustering and Classification**
  - Clustering of users help to discover groups of users with similar navigation patterns => provide personalized Web content
  - Clustering of pages help to discover groups of pages having related content => search engine
  - E.g. clients who often access webminer software products tend to be from educational institutions.
  - clients who placed an online order for software tend to be students in the 20-25 age group and live in the United States.
  - 75% of clients who download software and visit between 7:00 and 11:00 pm on weekend are engineering students

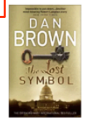
Pattern Discovery Tasks

- **Sequential Patterns:**
  - extract frequently occurring intersession patterns such that the presence of a set of items followed by another item in time order
  - Used to predict future user visit patterns=>placing ads or recommendations

- **Association Rules:**
  - Discover correlations among pages accessed together by a client
  - Help the restructure of Web site
  - Develop e-commerce marketing strategies - Grocery Mart

Frequently bought together

+



+

Inferno	The Lost Symbol	Angels And Demons
4.5★ (3,365) Assured	4.5★ (2,073)	4.6★ (2,123) Assured
₹226 ~~₹399~~ 43% off	₹343	₹226

1 Item ₹226 + 2 Add-ons ₹569 = Total ₹795

ADD 3 ITEMS TO CART

Customers who viewed this item also viewed

Yonex NANORAY Series Badminton Racket with a Half-length Cover
3.5 48
$39.99 - $80.90

Senston N80 Graphite Single High-grade Badminton Racquet, Professional Carbon...
4.5 137
$31.34 prime

Yonex Voltric 2017 New (7 NEO / 5FX / LITE) Badminton Racket (Racquet) 4U/G5 Strung...
4.0 23
$92.82 - $125.00

Yonex Nanoray 20 Badminton Racket 2016 NR20 Racquet 3U5G
4.0 44
$77.99 prime

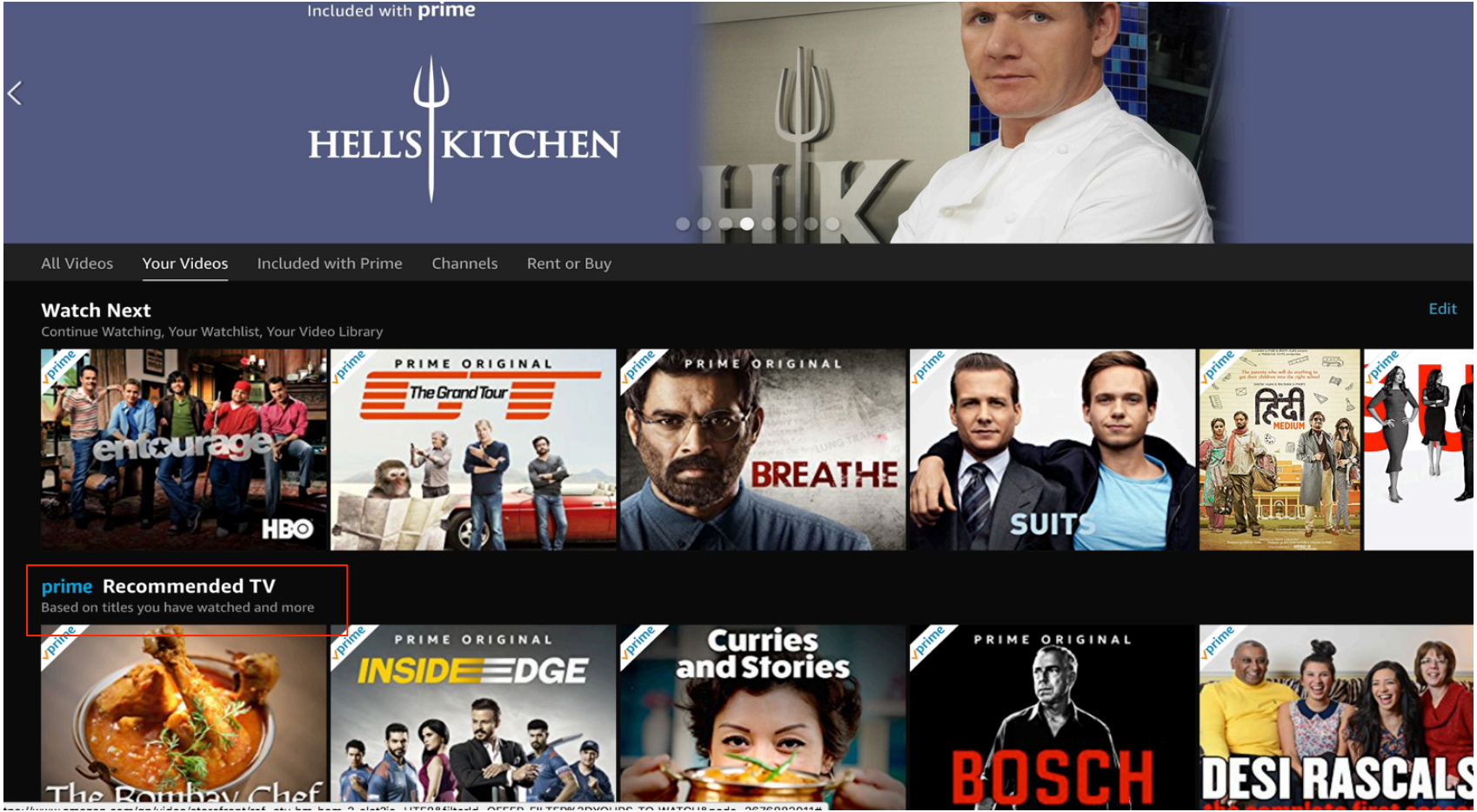
https://www.flipkart.com/ amazon.com

Pattern Analysis Tasks

- Pattern Analysis is the final stage of WUM, which involves the validation and interpretation of the mined pattern

- **Validation**:
  - to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process

- **Interpretation**:
  - the output of mining algorithms is mainly in mathematic form and not suitable for direct human interpretations



https://www.amazon.com/Amazon-Video

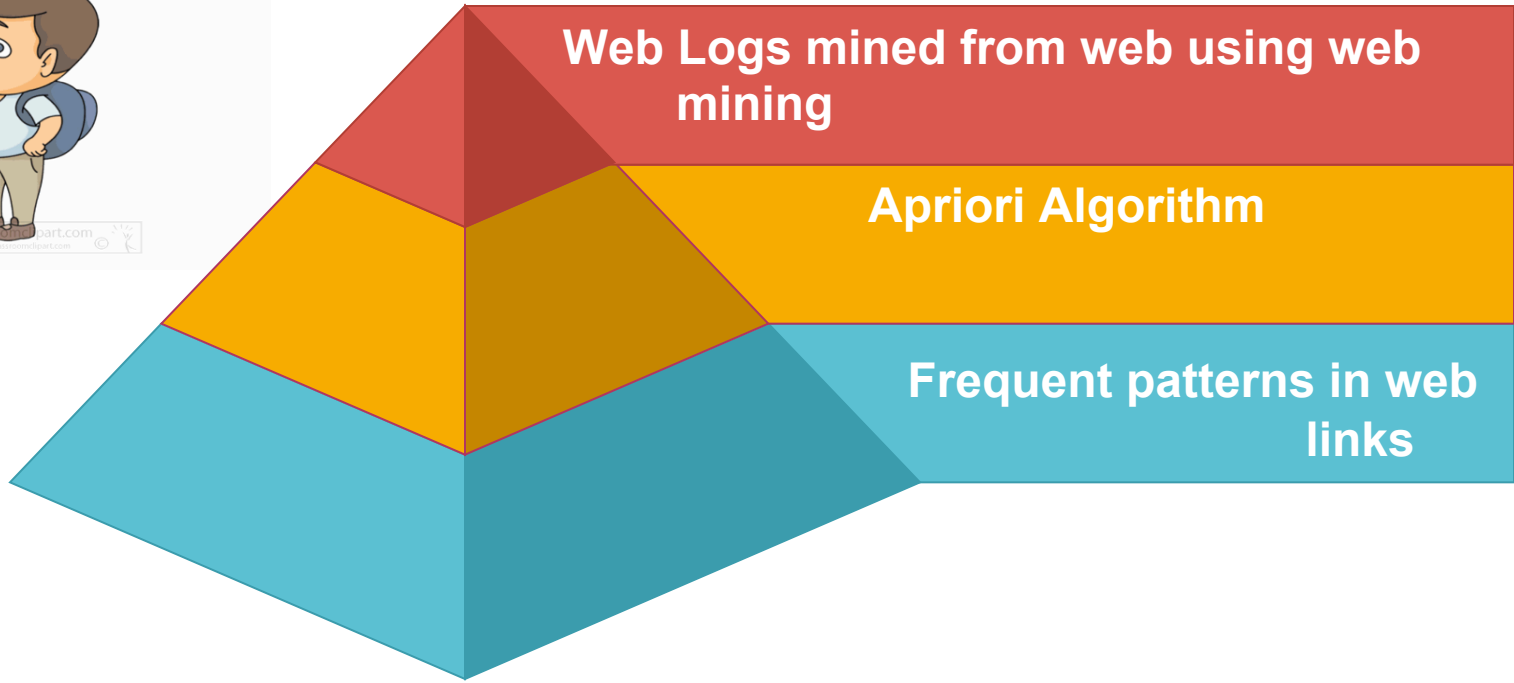
Summary of Research paper “Frequent Pattern Mining in Web Log Data using Apriori Algorithm”

Authors : S.VijayaKumar, A.S.Kumaresan, U.Jayalakshmi

International Journal of Emerging Engineering Research and Technology (IJEERT)

Volume 3, Issue 10, October 2015

Key Concepts Used



Step by step derivation of interesting rules using Apriori Algorithm on web logs

1

WEB LOG DATA

IP Address	URL Accessed	Time
rres1,ne.wi.ac.uk	/api/java.io.Bufferedwriter,html	01/Jan/1999
	/api/java.util.zip, CRc32.html	01/Jan/1999
	/api/java.io.Bufferedwriter,html	02/Feb/1999
	/java-tutorial/ui/animLoop.html	04/Feb/1999
	/atm /logiciels,html\	18/Feb/1999
	/relnotes/deprecatedlist.html	18/Feb/1999
Acasun.cekerd.edu	/perl/perlre.html	11/Jan/1999
	/java.tutorial/animLoop.html	12/Jan/1999
	/html4,0/struct/global,html	29/Jan/1999
	/api/java.util.zip,CRC32,html	29/Jan/1999
	/postgres/html-manual/query.html	29/Jan/1999
Acccs.francomedia.gc.ca	/java.tutorial/animLoop.html	05/Jan/1999
	/apache/manual/misc/API.html	05/Jan/1999
	/postgres/html-manual/query.html	05/Jan/1999
	/perl/perlre.html	12/Feb/1999
	/api/java.io.Bufferedwriter,html	12/Feb/1999
ach3.pharma.mcgill.ca	api/java.io.Bufferedwriter,html	06/Feb/1999
	/java-tutorial/ui/animLoop.html	06/Feb/1999
	/html4,0/struct/global,html	07/Feb/1999
	/postgres/html-manual/query.html	07/Feb/1999
	/relnotes/deprecatedlist.html	08/Feb/1999

Minimum Support count = 2
Minimum confidence = 60%

http://www.ijeert.org/pdf/v3-i10/9.pdf

2 Numbering URLs

URL	No.
/api/java.io.Bufferedwriter,html	1
/api/java.util.zip, CRc32.html	2
/java-tutorial/ui/animLoop.html	3
/atm /logiciels,html\	4
/relnotes/deprecatedlist.html	5
/perl/perlre.html	6
/html4,0/struct/global,html	7
/postgres/html-manual/query.html	8
/apache/manual/misc/API.html	9

3 Summary of Web log Data

Ip Address	URL Accessed
A	1,2,1,3,4,5
B	6,3,7,2,8
C	3,9,8,6,1
D	1,3,7,8,5

4 Scan database for count of each candidate

URL	Support Count
1	4
2	2
3	4
4	1
5	2
6	2
7	2
8	3
9	1

5 Calculation of L1

URL	Support Count
1	4
2	2
3	4
5	2
6	2
7	2
8	3

6 Calculation of L2

URL	Support Count
1,3	3
1,5	2
1,8	2
2,3	2
3,5	2
3,6	2
3.7	2
3,8	3
6,8	2
7,8	2

http://www.ijeert.org/pdf/v3-i10/9.pdf

7 Calculation of L3

URL	Support Count
1,3,5	2
1,3,8	2
3,6,8	2
3,2,8	2

8.1 Calculation of Confidence for L2

URL	Total Occurrences of X&Y	Total occurrences of X	Confidence
1,3	3	4	0.75
1,5	2	4	0.5
1,8	2	4	0.5
2,3	2	2	1
3,5	2	4	0.5
3,6	2	4	0.5
3.7	2	4	0.5
3,8	3	4	0.75
6,8	2	2	1
7,8	2	2	1

8.2 Calculation of Confidence for L3

URL	Total Occurrences of X&Y	Total occurrences of X	Confidence
<1,3>,5	2	3	0.67
<1,3>,8	2	3	0.67
<3,6>,8	2	2	1
<3,2>,8	2	2	1

http://www.ijeert.org/pdf/v3-i10/9.pdf

Conclusion of Research paper

It introduced the process of web log mining, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user’s navigation behavior.


Thank you