


Web data clustering
Current research status & trends

Prague, May 13th, 2004


by Athena Vakali
 Dpt. of Informatics, Aristotle University,
 Thessaloniki, Greece



Presentation content


- Clustering on the Web : Basics
- Web logs and user sessions identification
- User sessions clustering
 - Similarity-based
 - Model-based or probabilistic
- Web documents clustering
 - Compound documents
 - Logical info units
- The EDBT04 ClustWeb Workshop

Web clustering, by A. Vakali, Prague, May 13th 2004 2



Clustering on the Web : Basics


Web clustering, by A. Vakali, Prague, May 13th 2004 3



Some questions ...

- What is Web data clustering ?
- Why clustering on the Web ?
- What do we mean by "Web data" ?
- What type of clustering to apply ?
- What are the benefits ?


Web clustering, by A. Vakali, Prague, May 13th 2004 4



Web data clustering - Basics

- Organize data circulated over the Web into **groups / collections** in order to facilitate data availability & accessing, and at the same time meet user preferences
- The initial idea was to define the correlation distance / similarity measure between any two "elements"
 - Euclidean distance, Manhattan distance, cosine distance etc.

Web clustering, by A. Vakali, Prague, May 13th 2004 5



What is Web Data Clustering?

- Grouping Web objects into "classes" so that similar objects are in the same class and dissimilar Web objects are in different classes
- Discover distribution patterns and relationships between data attributes
- Employ Unsupervised learning

Web clustering, by A. Vakali, Prague, May 13th 2004 6



Why Clustering on the Web? some benefits ..

- *Increasing* Web information accessibility
- *Decreasing* lengths in Web navigation pathways
- *Improving* Web users requests servicing
- *Improving* information retrieval
- *Improving* content delivery on the Web
- *Understanding* users' navigation behavior
- *Integrating* various data representation standards
- *Extending* current Web information organizational practices



Types of Clustering on the Web

- Hierarchical clustering
- Partitional clustering
- Probabilistic clustering
- Graph-based clustering
- Fuzzy clustering
- Neural Network based clustering
- Hybrid approaches



Clustering practices

- Web server Log files
- Sessions identification
- Users clustering/grouping
- Pages clustering/grouping



Web logs & user sessions



What do we consider as "Web data" ?

- Web documents
 - A collection of Web Pages (set of related Web resources, such as HTML files, XML files, images, applets, multimedia resources etc.)
- Users' navigation sessions
 - The group of activities performed by a user from the moment the user enters a Web site to the moment the same user leaves it
 - The records of users' actions within a Web site are stored in a log file (each record in the log file contains the client's IP address, the date and time the request is received, the requested object and some additional information -such as protocol of request, size of the object etc.)



Users Session Identification

- Identification of unique users
 - Users with the same client IP are identical
- Identification of sessions
 - A new session is created when a new IP address is encountered or if the visiting page time exceeds a time threshold (i.e. 30 minutes) for the same IP-address



Web logs and definition of users' navigation sessions

- **Web Server Log File:** A Web log file is a collection of records of user requests for documents on a Web site, an example:

```
216.239.46.60 - - [04/Jan/2003:14:56:50 +0200] "GET
/-lpis/curriculum/C+Unix/Ergastiria/Week-7/filetype.c.txt HTTP/1.0" 304 -
216.239.46.100 - - [04/Jan/2003:14:57:33 +0200] "GET /-oswinds/top.html
HTTP/1.0" 200 869
64.68.82.70 - - [04/Jan/2003:14:58:25 +0200] "GET /-lpis/systems/r-
device/r_device_examples.html HTTP/1.0" 200 16792
216.239.46.133 - - [04/Jan/2003:14:58:27 +0200] "GET /-lpis/publications/crc-
chapter1.html HTTP/1.0" 304 -
209.237.238.161 - - [04/Jan/2003:14:59:11 +0200] "GET /robots.txt HTTP/1.0" 404
276
209.237.238.161 - - [04/Jan/2003:14:59:12 +0200] "GET /teachers/pitaas1.html
HTTP/1.0" 404 286
216.239.46.43 - - [04/Jan/2003:14:59:45 +0200] "GET /-oswinds/publication
```

<http://www.csd.auth.gr>



Problems with the Web logs processing

- not adequate/detailed info is provided
- there is no info about the content of the pages visited
- too many log records due to the visiting of image files, etc
- incomplete log recording due to the request servicing by proxies



Some Practices (I) : Data Cleaning

- **Data Cleaning:** removes log entries that are not needed for the mining process
 - e.g. Images, css files etc
- Typically, log entries are filtered:
 - Log entries with filename suffixes such as gif, jpeg, jpg
 - The page requests made by the automated agents and spider programs
 - The log entries that have a status code of 400 and 500 series
 - POST data (i.e. CGI request)



Some Practices (II) : Page Visiting Time Evaluation

- Why evaluating visiting page time?
 - The time spent on a page is a good measure of the user's interest in that page, providing an implicit rating for that page
- Page Visiting Duration : Time difference between consecutive page requests
- **Drawback :** some users are left to a page because they have completed a search and they no longer wish to navigate



Some Practices (II) : Heuristics

- **Session Identification:** groups users' page references into user sessions based on some heuristics :
 - Heuristics based on IP, and session time-outs (i.e. 30 minutes) used to identify unique user sessions
 - Intra-session transactions can be obtained based on a model of user behavior (involves classifying references as "content" or "navigational" for each user)
 - Weights are assigned to each Web page based on some measures of user interest (e.g., duration of viewing a Web page)



User sessions clustering

Why clustering users navigation sessions ?

Clustering users' navigation sessions : Groups together a set of users' navigation sessions having similar characteristics

User Grouping

- Discover groups of users exhibiting similar browsing patterns

Web Page Grouping

- Discover groups of pages having related content
- based on how often URL references occur together across user sessions

benefits

Web clustering, by A. Vakali, Prague, May 13th 2004 19

Clustering users' navigation sessions – benefits (1)

- Users grouping helps to discover groups of users with similar navigation patterns
 - Provide personalized Web content
- Web personalization
 - Any action that adapts the information or services provided by a Web site to the needs of a particular user (or a set of users)
- Benefits
 - discover the preference and needs of individual Web users in order to provide personalized Web site for certain types of users
 - examine general user navigation patterns in order to understand how general users use the site

Web clustering, by A. Vakali, Prague, May 13th 2004 20

Clustering users' navigation sessions – benefits (2)

- Provide information about :
 - What are the set of pages frequently accessed together by Web users? (*frequent itemsets*)
 - What page will be fetched next? (*association rules*)
 - What are the paths frequently traversed by Web users? (*sequential patterns*)
- Clustering Web users' sessions are useful:
 - To improve Web site design
 - To develop prefetching and Web caching policies
 - To recommend related pages
 - To collect business info about Web users behavior

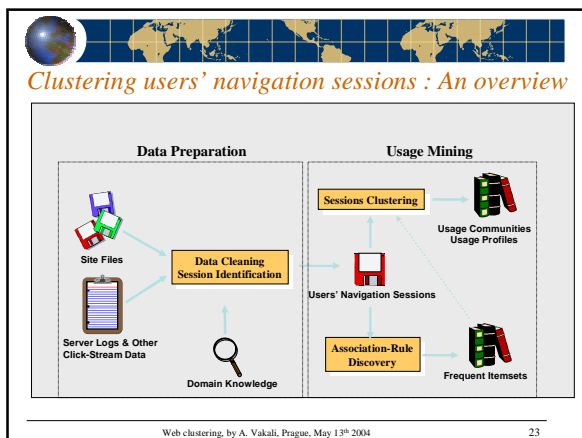
Applications : e-commerce, e-learning, e-Gov

Web clustering, by A. Vakali, Prague, May 13th 2004 21

A generic Clustering approach for users' navigation sessions

1. Determine the attributes to be used to estimate similarity between users' sessions, or determine the **users' session representation**
2. Determine the "strength" of the relationships between the attributes, or a **similarity measure (correlation distance)**
3. Apply **clustering algorithms** to determine the classes/clusters to which each user session will be assigned

Web clustering, by A. Vakali, Prague, May 13th 2004 22



Algorithms for Sessions Clustering

- **Similarity-based clustering**
 - Parameters: distance functions, number k of clusters
 - Approaches
 - Hierarchical - e.g. determine a hierarchy of clustering, merging always the most similar clusters
 - Partitional - Determine a "flat" clustering into k clusters (with minimal costs)
- **Model-based or Probabilistic clustering**
 - Parameters: number k of clusters
 - Determine a probability model for each cluster

Web clustering, by A. Vakali, Prague, May 13th 2004 24



Similarity-based session Clustering (I)

- **Originally**, sessions clustering efforts considered sessions as unordered sets of "clicks", where the number of common pages visited was a similarity indication between sessions (measures used: Euclidean dist., cosine measure, Jaccard coefficient etc).
- **Later on**, it was recognized that the order of visiting pages is important, since for example visiting a page A after a page B is not the same information as knowing that Both A and B belong to the same session. In this context, we have the:
- **Sequence Alignment Method (SAM)** [Hay01, Wan02], where *sessions are chronologically ordered sequences of page accesses*.
 - SAM measures similarities between sessions, taking into account the sequential order of elements in a session.
 - Define: Web pages similarity (based on the URL "token") and then, sessions similarity (dynamic programming method to match related sessions - scoring function). SAM distance measure between two sessions is defined as the number of operations that are required in order to equalize the sessions.



Similarity-based session Clustering (II)

- **Clickstream analysis** (recent work in [Kot03]) - how to evaluate similarities between two clickstreams?
 - edit (or Levenhstein) distance: cost of transformations that result in two clickstreams to be identical
 - LCS (Largest Common Subsequence): length of the largest subsequence common between two clickstreams
 - Similarity between 2 clickstreams requires finding similarity/distance between 2 page views. Since semantic analysis is not possible, the degree of similarity between two page views is proportional to their relative frequency of cooccurrence.
- **Generalization-based clustering** [Fu99]
 - Uses page URLs to construct a hierarchy, for categorizing the pages (partial ordering of Web pages, leaf is the Web page file, non-leaf nodes are the *general pages*).
 - *Then*, the pages in each user session are replaced by the corresponding general pages and clustered using the BIRCH algorithm.



Probabilistic session Clustering (I)

- Assume a model for each cluster (the number of cluster is pre-determined) and find best fit of models to data using the Expectation-Maximization (EM) algorithm (originally EM proposed by [Dem77])
- Each cluster is modeled by a finite-state Markov model with a number of parameters:
 - Markov models popular to characterize probability of referencing page j after page i
 - e.g. First-order Markov Models or higher order Markov Models, HMMs (Hidden Markov Models) etc. [And02, Bal03, Cad03, Des01, Sar00, Sen03, Smy99]
- The number of clusters is determined by using:
 - BIC (Bayesian Information Criterion)
 - Bootstrap methods or cross-validated likelihood using re-sampling ideas
 - Bayesian approximations



Model-based (Probabilistic) Clustering

- **Model-based (or Probabilistic) Clustering Methods**
 - **optimize** the "fit" between the given data and a mathematical model
 - **based** on the assumption that the data are generated from a probability distribution
- **Model-based (Probabilistic) clustering problem**
 - Find the model structure
 - Find the model parameters for the structure that best fit the data




Generative Mixture-Based Cluster Model

- Draw an individual i from the overall population.
- The individual is assigned to one of K clusters $1 \leq k \leq K$, with probability $p(c_i=k)$, $\sum_{k=1}^K p(c_i=k) = 1$ where c_i indicates the cluster membership.
- Each cluster k , $1 \leq k \leq K$, has a data generating model $p_k(D_i/\Theta_k)$ where Θ_k are the parameters of p_k
- D_i now generated for an individual by $p_k(D_i/\Theta_k)$ once cluster membership $c_i=k$ is known and given Θ_k .



Model-based(Probabilistic) Clustering over Web Log Data

- Clustering user sessions according to the amount of time spent on common pages
- When a user arrives at a Web site, his/her session is assigned to one of the clusters with some probability
- Given that a user's session is in a cluster, his/her next request in that session is generated according to a probability distribution specific to that cluster




Web browsing example (I)

User 1	Session 1	2	3	2	2	3	3	1	1	3	1	3	1	3
	Session 2	3	3	3	1	1								
User 2	Session 1	7	7	7	7	7	7	7						
User 3	Session 1	1	5	1	1	5	1	1	1	1	1	1	1	1
	Session 2	5	1	1	5									
	Session 3	1	3	3	1	5	1	1						

- Each individual has a set $D_i = \{s_1, s_2, \dots, s_{n_i}\}$ where each s is a sequence that represent the observed record of page requests for individual i and the different sequences represent the different sessions.

Web clustering, by A. Vakali, Prague, May 13th 2004 31




EM-Based Clustering Algorithm for Clustering Individuals [Cad03]

- Consider N individuals each having a data set D_i . Let each D_i consist of n_i observations d_{ij} . Each d_{ij} represents another smaller data subset.
- According to the generative cluster model each i has a pdf: where $\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$

$$p(D_i/c_i, \Theta) = p(D_i/\Theta_{c_i})$$
- and each c_i is the cluster identity of the i th individual
- Assuming that the observations are conditionally independent, the prob. that i belongs to c_i is

$$p(D_i/c_i, \Theta) = \prod_{j=1}^{n_i} p(d_{ij}/\Theta_{c_i})$$


Web clustering, by A. Vakali, Prague, May 13th 2004 32



Web Browsing example (II)


- Given the definition of the likelihood function, the EM procedure becomes:
- repeat
 - E step:** a straight-forward evaluation of class conditional probability for each individual under each of the K cluster models using values of parameters Θ .
 - M step:** update parameters Θ to obtain maximum likelihood or maximum a posteriori parameter (MAP)
- until a condition is satisfied
 - // e.g. a convergence criterion like difference between two successive values of MAP

Web clustering, by A. Vakali, Prague, May 13th 2004 33



Web documents clustering


Web clustering, by A. Vakali, Prague, May 13th 2004 34



Document-oriented approaches for Clustering on the Web

- Clustering of Web documents helps to discover groups of pages having related content
- Web communities**
 - A set of Web pages that link to more Web pages in the community than to pages outside of the community
 - A web community enables web crawlers to effectively focus on narrow but topically related subsets of the web.
- Logical document**
 - A set of Web pages with similar content
- Benefits**
 - Improves Web information retrieval (e.g. search engines)
 - Improves content delivery on the Web

Web clustering, by A. Vakali, Prague, May 13th 2004 35



Compound Documents (I)

- Techniques used to recognize and group hypertext nodes into cohesive documents can improve information retrieval results.
- Compound documents:** A compound document is a set of URLs that contains at least a tree embedded within the document. Most compound document hyperlink graphs are either strongly connected or nearly so.
- Necessary condition** for a set of URLs to form a compound document : their link graph should contain a vertex that has a path to every other part of the document

Web clustering, by A. Vakali, Prague, May 13th 2004 36



Compound Documents structure

- Compound documents are commonly found to contain at least one of the following graph structures within their hyperlink graph
 - **linear paths:** There is a single ordered path through the document, and navigation to other parts of the document are usually secondary (e.g. news sites with next link at the bottom)
 - **Fully connected:** These types of documents have on each page, links to all other pages of the document (e.g. short technical docs and presentations)
 - **Wheel documents:** They contain a table of contents and have links from this single table of contents to the individual sections of the document (toc is a kind of "hub" for the document)
 - **Multi-level documents:** Complex documents that may contain irregular link structures such as multilevel table of contents



Heuristics for compound documents identification

- **Hierarchical structure of URLs**
They reflect the intention of page authors and they can lead to potential compound documents
- **Page contents**
This approach analyzes the contents of pages in order to detect logical information units
- **Link structure**
Graph theoretic properties can clue to detecting strongly related pages



Logical Info Units on the Web

- A set of Web pages with similar content.
- Then, a data unit for the Web data retrieval should not be a page but a connected subgraph corresponding to one **logical document**
- Introduce the concept of route links [Taj99], then rank minimal subgraphs under a given query and consider distribution of query keywords within subgraphs



The idea of Web Communities



Web communities were proposed [Gre04] on the basis of the evolution of an initial set of hubs and authoritative pages, such that the behavior of users is captured with respect to the popularity of existing pages for the topic of interest



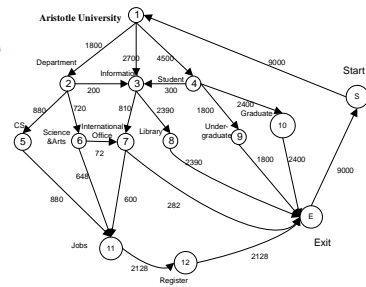
identifying collections of web communities

- Hyperlink Induced Topic Search Algorithm (HITS - Google's Approach)
- Graph cuts and partitions
- Maximum Flow and Minimal Cuts
- PageRank algorithm
- Other methods :
 - **Bibliometric methods:** They define a notion of similarity for pages that do not directly link to one another
 - **Bipartite cores:** They consist of pages that have high bibliographic metrics with respect to each other



Clustering web pages – A probabilistic based approach (I)

Link Graph
[Zhu02]:
an example





Clustering web pages – A probabilistic based approach (II) building Markov Models from Web Log Files

- Markov model: $\langle S, Q, L \rangle$
 - S - state space containing all the nodes in link graph
 - Q - matrix of 1-step transition probabilities between nodes
 - L - initial probability distribution on the states in S
 - m-order Markov chain
 - the next page is dependent only on the last visited m pages
 - m-order n-step Markov chain
 - the n-th page to be visited in the future is dependent only on the last visited m pages



Clustering web pages – A probabilistic based approach (III)

Using Markov Models for Link Prediction

- Link prediction is based on a m-order n-step Markov chain
 - given visited m pages, calculate the probability of visiting page a within the next n steps
 - the probability: weighted sum of probabilities of visiting page a at 1st to n-th step



ClusWeb Workshop - EDBT 2004



EDBT Workshop Chairs : J. Pokorny – A. Vakali Contribution (I)

- Web sources clustering
 - Web sources structured under a schema (XML-oriented)
 - Schemas define the *object domain* of a source (e.g., Books, Movies) and its *query capabilities*
 - Goal: Clustering Web sources by their schemas (e.g. attributes in query interfaces)
- Similarities and clustering framework definition
 - Tree structural similarities (e.g. tree edit distances)
 - Contextual similarities of Web pages based on the hyperlink structures
 - Pattern modeling framework definition



EDBT Workshop Contribution (II)

- Web logs - selective clustering & dissemination
 - Selective dissemination of information
 - Clustering Web query logs for flexible and productive query systems
 - Representing Web sources with feature spaces
- Documents-based clustering
 - Extracting knowledge from the content of the Web document (the location and frequency of words occurring in a Web document)



Our Research Group Focus on...

- Clustering Web users Sessions
 - Develop a probabilistic validation algorithm for Web users' clusters
 - Develop a novel probabilistic clustering algorithm for Web users' sessions
- Web prefetching
 - Identify "Web Page Communities" (using graph-based clustering algorithms) from a popular Web site
 - Clustering Web documents using XML standard



Some Indicative Publications...

- A. Vakali, G. Pallis: "Content Delivery Networks: Current Status and Trends", IEEE Internet Computing, Vol. 7, No. 6, pp.68-74, 2003
- G. Pallis, K. Stoupa, A. Vakali: Storage and Access Control Issues for XML documents, book chapter in the book "Web Information Systems", editors: D. Taniar and W. Rahayu, Idea-Group Publishing, USA, 2003.
- A. Vakali, E. Terzi, E. Bertino, A. Elmogamrid, "Hierarchical Data Placement for Navigational Multimedia Applications", Data and Knowledge Engineering Journal, Elsevier, to appear.
- A. Vakali: "A Simulation Model for Prefetching and Caching in a Parallel Storage Subsystem", International Journal of Modeling and Simulation, accepted to appear 2002.
- A. Vakali, E. Terzi, L. Angelis, M.-S. Hacid: "Multimedia Documents Storage: An Evolutionary based Application", Proceedings of the International Workshop on Multimedia Data and Document Engineering, in conjunction with the 7th International Conference Reverse engineering Technologies for Information Systems (ReTIS), pp. 36-44, Lyon, France July 2001.



OSWINDS Research Group

Website: <http://www.csd.auth.gr/~oswinds>

Faculty members

- Athena Vakali, Ph.D., Assistant Professor
- Lefteris Angelis, Ph.D., Assistant Professor

PhD Students

- Vassiliki Koutsonikola
- Lefteris Moisiadis
- George Pallis
- Konstantina Stoupa
- Theodosios Theodosiou



Presentation References-Clustering users' navigation sessions (1)

- [And02] C. R. Anderson, P. Domingos, D. Weld: Relational Markov Models and their Application to Adaptive Web Navigation. Proceedings 8th International Conference on Knowledge Discovery and Data Mining, pp.143-152, ACM, New York, 2002.
- [Bal03] Baldi, P., Frasconi, P. Smyth: Modeling the Internet and the Web, Wiley 2003.
- [Bon01] A. Bonerjee, and J. Ghosh: Clickstream Clustering using Weighted Longest Common Subsequences, Proceedings of the Workshop on Web Mining, SIAM Conference on Data Mining, pp.33 - 40, Chicago IL, April 2001.
- [Cad03] I. V. Cadez, D. Heckerman, C. Meak, P. Smyth, S. White: Model-based clustering and visualization of navigation patterns on a Web site, Journal of Data Mining and Knowledge Discovery, in press. Extended version of ACM SIGKDD 2003.
- [Cho03] S. Chakrabarti: Mining the Web, Morgan Kaufmann Publishers, 2003.
- [Che03] Chen, A. Wei-Ches Fu, F. Chi-Hung Teng: Optimal Algorithms for Finding User Access Sessions from Very Large Web Logs, World Wide Web: Internet and Information Systems, Vol. 6, 259-279, 2003.
- [Dem77] A. P. Dempster, N. M. Laird, D. B. Rubin: Maximum Likelihood from Incomplete Data via the EM Algorithm, J. R. Statistics Society B 39, pp. 1-22, 1977.
- [Des01] M. Deshpande, and G. Karayiannis: Selective Markov Models for Predicting Web Page Accesses, Proceedings of SIAM Conference Data Mining SIAM Press.
- [Fre98] C. Fraley and A. Raftery: How Many Clusters? Which Clustering Method? Answers via model-based cluster analysis. Computer Journal, Vol. 41, pp. 578-598, 1998.
- [Fu99] Y. Fu, K. Sandhu, M.-Y. Shih: Clustering of Web users based on access patterns, WEBKDD99, 1999.
- [Har02] D. Hand, H. Mannila, P. Smyth: Principles of Data Mining, MIT Press, 2002.
- [Hay01] B. Hay, K. Vanhoof, G. Wets: Clustering navigation patterns on a website using a sequence alignment method, Proceedings of 17th International Joint Conference on Artificial Intelligence, August 4, Seattle, Wash., USA, 2001.
- [Huo01] Z. Huang, J. Ng, D. W. Cheung, W.K. Ng and W.K. Ching: A Cube Model for Web Access Sessions and Cluster Analysis, Proceedings of WEBKDD 2001 Third International Workshop, August 26, 2001, San Francisco, CA, USA.
- [Jag99] A.K. Jain, M.N. Murty, P.J. Flynn: Data Clustering: A Review, ACM Computing Surveys, Vol. 31, No. 3, September 1999.
- [Kar] G. Karayiannis: METIS: Family of Multilevel Partitioning Algorithms: <http://www-users.cs.umn.edu/~karayiannis/metis/>.
- [Kor03] R. Kothari, P. A. Mittal, V. Jain, M. K. Mohana: On using Page Cooccurrences for Computing Clickstream Similarity, Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, USA, May 1-3, 2003.



Presentation References-Clustering users' navigation sessions (2)

- [Lev03] M. Levene and G. Loizou Computing the entropy of user navigation in the web. To appear in International Journal of Information Technology and Decision Making, Vol. 2, pp. 459-476, 2003.
- [Ng02] R. T. Ng, J. Han: CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 1, pp. 1003-1016, Jan/Feb. 2002.
- [Sar00] R. R. Sarukkai: Link Prediction and Path Analysis using Markov Chains. Computer Networks 33, pp. 377-386, 2000.
- [Sen03] R. Sen, and M. H. Hansen: Predicting a Web user's next request based on log data, Journal of Computations Graph Statistics, 2003.
- [Sho97] C. Shahabi, A. M. Zarkesh, J. Adibi, and V. Shah: Knowledge Discovery from Users Web Page Navigation, IEEE RIDP 97, 1997.
- [Smy99] P. Smyth: Probabilistic model-based clustering of multivariate and sequential data, Proceedings of the Seventh International Workshop on AI and Statistics, Jan. 1999.
- [Su01] J. Z. Su, Q. Yang, H. H. Zhang, X. Xu, Y. Hu: Correlation-based Document Clustering using Web Logs, Proceedings of 34th Annual Hawaii International Conference on System Sciences (HICSS-34), Maui, Hawaii, Jan. 3-6, 2001.
- [War02] W. Wang, O. R. Zaiane: Clustering Web Sessions by Sequence Alignment, Proceedings of 13th International Workshop on Database and Expert Systems Applications (DEXA 2002), Aix-en-Provence, France, 2-6 Sep. 2002.
- [Xiao] J. Xiao, and Y. Zhang: Clustering of Web Users Using Session-based Similarity Measures, IEEE, pp. 223-228, 2001.
- [Yan01] Q. Yang, H. H. Zhang, T. T. Y. Li: Mining web logs for prediction models in WWW caching and prefetching, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, August 26-29, 2001, San Francisco, CA, USA.
- [Ypm02] A. Ypma, J. Heskes: Categorization of web pages and user clustering with mixtures of hidden Markov models. In Proceedings of WEBKDD 02, pp. 31-45, 2002.
- [Xie01] Y. Xie, V. Phoha: Web user clustering from access log using belief function, Proceedings of the ACM K-CAP01, First International Conference on Knowledge Capture, pp. 202-208, Victoria, British Columbia, Canada, Oct. 22-23, 2001.
- [Zha96] T. Zhang, R. Ramakrishnan, M. Livny: BIRCH: An Efficient Data Clustering Method for Very Large Databases, Proceedings 1996 ACM-SIGMOD International Conference Management of Data, pp. 103-114, Montreal, Canada, June, 1996.



Presentation References- Web Documents clustering

- [Eir03] N. Eiron, K. S. McCurley: Untangling Compound Documents on the Web, Proceedings of the ACM Hypertext, pp. 85-94, 2003.
- [Flo03] G. W. Flake, K. Tsoutsoulis, L. Zhukov: Methods for Mining Web Communities: Bibliometric, Spectral, and Flow, Overture Research Technical Report OR-2003-004.
- [Flo23] W. Flake, S. Lawrence, C. Lee Giles, Frans Goetzac: Self-Organization and Identification of Web Communities, IEEE Computer, Vol. 35, No. 3, pp. 66-71, 2002.
- [Flo00] G. W. Flake, S. Lawrence, C. Lee Giles: Efficient identification of Web communities, Proceedings of the 6th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 150-160, Boston, Massachusetts, United States, 2000.
- [Gib03] D. Gibson: Analyzing Repetition in Web Communities, Proceedings of the International Conference on Internet Computing, IC 03, Las Vegas, Nevada, USA, June 23-26, 2003 [Gib04].
- [Gred04] G. Greco, S. Greco, E. Zuppano: Web Communities: Models and Algorithms, World Wide Web, Volume 7, Number 1, pp. 58-82, Mar. 2004.
- [Huo02] J. Hou, Y. Zhang: Constructing good quality web page communities, Proceedings of the 13th Australasian conference on database technologies, January 2002.
- [Lar99] E. de Lara, D. S. Wallach, and W. Zwanevoo: A Characterization of Compound Documents on the Web, Technical Report 18-99-251, Department of Computer Science, Rice University, November 1999.
- [Taj99] K. Tajima, K. Hatano, T. Matsukura, R. Sano, K. Tanaka: Discovery and retrieval of logical information units in Web, Proceedings of the Workshop on Organizing Web Space (WOWS 99), in conjunction with ACM DL, pp. 13-23, Berkeley, CA, August 1999.
- [Zhu02] J. Zhu, J. Han and J. G. Hughes: Using Markov chains for Link Prediction in Adaptive Web Sites, Software 2002, Springer Verlag, LNCS 2311, pp. 60-73, 2002.