published as: Christoph Meinel, Long Wang: Building Content Clusters based on Modelling Page Pairs;
Proceedings of International Conference on Frontiers of WWW Research and Development (APWeb 2006);
Harbin (China), pp.856-861, Springer LNCS3841.

Building Content Clusters Based on Modelling Page Pairs

Christoph Meinel and Long Wang

Hasso Plattner Institut, Potsdam University, 14482 Potsdam, Germany
long.wang@hpi.uni-potsdam.de

Abstract. We give a new view on building content clusters from page pair models. We measure the heuristic importance within every two pages by computing the distance of their accessed positions in usage sessions. We also compare our page pair models with the classical pair models used in information theories and natural language processing, and give different evaluation methods to build the reasonable content communities. And we finally interpret the advantages and disadvantages of our models from detailed experiment results.

1 Introduction

In [19], Niesler used the distance between two words acting as a trigger-target pair to model the occurrence correlations within a word-category based language model. In this paper, we use "*heuristic importance*" to depict the importance of one page to attract visitors to access another page. The "*heuristic importance*" is measured by computing the distance of their access positions in usage sessions.

The methods to reconstruct sessions are classified into five different standards[15, 6, 21, 10, 11]. These five standards show the difference views on the binary relations of two accessed pages in reconstructing usage sessions. Web usage patterns are mostly defined on the association rules, sequential patterns and tree structure[2, 5, 6, 12]. Detailed definitions of different actions performed by visitors are given in [8, 9]. The binary relationship between every two pages in usage patterns are modelled on the co-occurrence happenings[2], time sequential[2, 5, 6] and structural characteristics [9, 12]. Clustering and classification of users are investigated in [1, 3, 21, 13, 18]. The binary relations between every two pages are computed on conditional possibilities or Markov chains [17], or on the content attributes[7]. We name the required terms in section 2 and give the general clustering method and web site modelling in section 3. We explain the evaluation measurements in section 4 and discuss our experiments in section 5. Section 6 is a short summary.

2 Problem Statements

A page pair is named as $Pair(p_{trigger}, p_{target})$. The heuristic importance is named as $Hr(p_{trigger}, p_{target})$. For a page p in a given session s, we use Pos_p to name the position of this page p in this session.

X. Zhou et al. (Eds.): APWeb 2006, LNCS 3841, pp. 856–861, 2006.

We improve this distance used in [19] to model the heuristic importance from trigger page to target page. Within a session s, the mutual relation from page $p_{trigger}$ to page p_{target} is named as: $M_s(p_{trigger}, p_{target}) = \frac{\sqrt{Pos_{p_{trigger}} Pos_{p_{target}}}}{|Pos_{p_{target}} - Pos_{p_{trigger}}|+1}$, where $Pos_{p_{trigger}}$ is the position for page $p_{trigger}$ in this session and $Pos_{p_{target}}$ for the page p_{target}.

Over the session set S, the heuristic importance from $p_{trigger}$ to p_{target} in $Pair(p_{trigger}, p_{target})$ is defined as:

$$Hr(p_{trigger}, p_{target}) = \frac{\sum_{n}^{i=1} M_{s_i}(p_{trigger}, p_{target})}{\sum_{i=1}^{n} |s_i|/2} : s_i \in S',$$

where *S'* is the sub set of S but includes all the sessions in which $p_{trigger}$ was accessed before p_{target}.

Here we also give the definitions of other methods to model page pairs.

Method 1 (SUP): A page pair is symbolized as two adjacently accessed pages in sessions. Given a session set S, $L(S)$ represents the corresponding binary page relation set, and *I(S)* the set reducing all the repeat happenings of the same binary relations in *L(S)*. The support of the page pair $PA = Pair(p_{trigger}, p_{target})$ in session set S is defined as: $supp_P = \frac{|\{PA_i | PA_i = PA, PA_i \in I(S)\}|}{|L(S)|}$. So we use this support as the heuristic importance of $p_{trigger}$ to p_{target}. This measurement is widely used in web usage mining [1, 2, 5, 6, 8, 12].

Method 2 (IS): A page pair is symbolized as two adjacently accessed pages in sessions. [18] used $Hr(p_{trigger}, p_{target}) = \frac{Pr(p_{target} p_{target})}{\sqrt{Pr(p_{trigger}) Pr(p_{target})}}$ to compute the heuristic importance of page $p_{trigger}$ to page p_{target}. This model is also used in computing the mutual information of in model natural language [14].

Method 3 (CS): The heuristic importance is characterized by the conditional possibility: $Hr(p_{trigger}, p_{target}) = Pr(p_{target} | p_{trigger})$. This measurement is also named as *confidence* in data mining and n-Markov chain is widely used in personalized recommendation and adaptive web sites [17].

3 Clustering Method and Site Modeling

The clustering method that finds the related page communities from page pairs is introduced in this section. An overview of the algorithm is given in follows:

```
Input: web server usage logs
Output: page clusters
1.Recover sessions from web usage logs,
2.Scan the recovered sessions and build
  page pairs by computing heuristic importance,
3.Create the graph from page pairs and find the cliques.
```

The method to recover sessions for different users has been detailed discussed in [11, 21], individual accessing behaviours are also recovered in this step for further interesting usage pattern mining [9]. In [13], clustering mining method was introduced in the PageGather system.

Table 1. Piece of Weights of Pages on HPI Site

URL	Weight
/	100
/lehre.html	16.7
/index.html	12.6
/support/sitetmap.html	3.33
/lehre/studienprojekte.html	2.08

Web has been modelled by many ways, most of which is based on the graph theory. PageRank[7] and HIT[4] are the two famous methods. Besides the graph model, role-based model was used in [6] and n-Markov was used in many personalized recommendations[17, 21].

We improved the method from HIT[4] to model the page relations within a web site. In this model, a web site is dedicated to several particular topics, and its semantic space can be formed based all the concepts related to these particular topics, and all the concepts are organized as a concept hierarchy. Each page within a web site is given a concrete numerical definition represented the corresponding sub set of concepts, and this numerical weight is computed by weight propagation step by step from the home page, which represents the whole concept set. We call the page that disperse its concept as "*host*", and the page that inherits concepts from "*host*" as "receiver". The weight of one concept w_{p_c} from a "*host*" is equally divided by all the "*receives*" that inherit this concept, and on the other hand, different concepts for a "*receiver*" is inherited from different "*host*".

Given a page p, its weight wp is computed as: $w_p = \sum_k^{i=1} p.w_{c_i}$ and $p.w_{c_i} = \frac{q.w_{c_i}}{n}$, where k is the number of different concepts for p, q is the host of p for concept c_i, and n is the number of receivers that inherit concept c_i from q.

During computing w_p for every page, the weight of a concept for a page is reduced with the weight propagation from the home page, so w_p represents the importance of its corresponding semantic value from the point web designer. The distribution and propagation of concept weights like our definition are universally observed in the frame work designing of many web sites. We illustrate this model on the content main frame of www.hpi.uni-potsdam.de. There are 67 different pages for the main frame, and they are organized as a tree structure. With the help of automated interface design, among these 67 pages, every two pages are directly connected. This helps greatly to reduce the affect of navigation hyperlink for pair analysis. The table 1 shows the weight for some pages.

4 Evaluation Measurements

The *number of clusters* and their *average size* are the two important measurable criteria for the success of a clustering method, and *distinctiveness* and *coverage* are the other two criteria for the quality of clusters. Given a set of M which is built by based on one clustering method, *distinctiveness* is given by the following equation:

$Distinctiveness(M) = \frac{|P'|}{\sum_{i=1}^{k}|P_i|}$, where P is the set of pages appearing as least in one clusters, and P_i is the pages used in *i-th* cluster, if there is k clusters in *M*. And *coverage* is given as: $Coverage(M) = \frac{|P'|}{|P|}$,where P' is the set of pages appearing at least in one cluster, and P is the pages that need to be clustered.

In our scenario, we add another two criteria: *semantic dependence* and *popularity. Semantic dependence* is defined as: $Semantic - dependence(M) = \frac{|C|}{\sum_{i=1}^{k}|C_i'|}$. In the above formula, C is the set of content categories that a web site belongs to, C_i is content categories that the $i-th$ cluster has, if there is k clusters in M. We use the average support of the page pairs appearing in at least one cluster to name the popularity of the clustering model, the popularity of a model M is: $Popularity(M) = \frac{\sum_{i=1}^{n}Pr(PA_i)}{|PA|}$. In this popularity formula, PA is the set of page pairs that appear in M, and $Pr(PA_i)$ is the possibility of page pair PA_i over the usage record set.

In [18], *Gold standard* is named as the expert criterion in general evaluation method that is used to find "ideal solution" to a problem. The methods to reduce the subjective bias from experts is trying to get as more suitable experts as possible. In web usage mining, the ideal evaluation for a content improving schema is the direct feedback from client sides. But in web applications such direct feedback is uncontrollable. We are pushed to raise three measurements to evaluate usage models:

1. `If similar patterns happen in different models,`
   `then this pattern is useful.`
2. `If similar patterns happen in different periods of time,`
   `then this pattern is valuable.`
3. `If a model reflects the changes of the content reorganization,`
   `then this model is reasonable.`

5 One Case Study

We take the content frame from www.hpi.uni-potsdam.de as the improving target, which includes 67 different pages with different URLs. We take two pieces of web logs for clustering page pairs, one is from 01.03.2005 to 31.03.2005, and the other is from 01.04.2005 to 30.04.2005. By fetching the related usage information of 67 target frame pages, we get 16314 sessions from March, and 18546 from April.

Here we show the validity of our method by one case study:

P_3=/lehre/vorlesungen.html, *P_4=/lehre/bachlor.html*, and
P_5=/lehre/master.html.

These three pages have the same semantic importance computed as in section 4, because they are linked from the same source page as */lehre.html*. This means that these three pages have no bias on the web designer's side. But based on page pairs modeled from usage data, these three pages have some clear bias on heuristic importance.

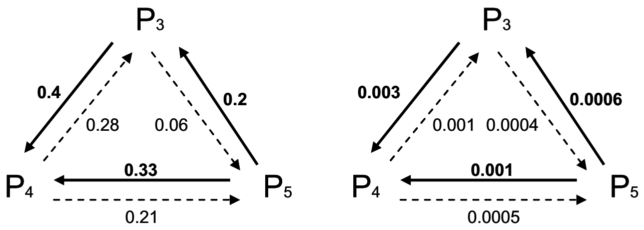

Fig. 1. Page Clusters base on CS and DS in March Logs

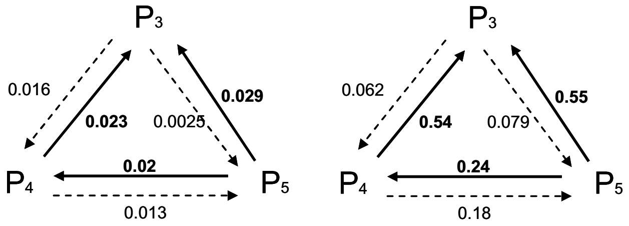

Fig. 2. Page Clusters base on CS and DS in April Logs

In the above tow figures, we discriminate the different directions of heuristic importance within a page pair by using different lines: the bold line means a higher heuristic importance and the dashed line means a lower heuristic importance within the same page pair. From the four page clusters in these two figures, we find P_5 has a higher heuristic importance to P_3 and P_4 than those from P_3 and P_4 to P_5, which happens in two different period of logs based on two different models. Based on task-oriented evaluating measurements in section 6, we can naturally conclude that $P_3 < -P_5- > P_4$ is a very useful page cluster and helps for improving content organization.

6 Conclusion

In this paper, we investigate the problem of building content clusters based on modeling page pairs by computing the position distance between source page and target page. Some questions are still open for further investigation, for example, measuring the difference between usage patterns and original web organization.

References

1. A. Banerjee, J. Ghosh: Clickstream Clustering using Weighted Longest Common Subsequences. In Workshop on Web Mining, SAIM, (2001).
2. B. Berendt, and M. Spiliopoulou: Analysis of navigation behaviour in web sites integrating multiple information systems. The VLDB Journal, (2000).

3. J. Heer, E. Chi: Mining the Structure of User Activity using Cluster Stability. In Workshop on Web Analytics, SIAM, (2002).
4. J. M. Kleinberg: The Web as a Graph: Measurements, Models, and Methods. In 5th International Conference on Computing and Combinatorics, (1999).
5. Jian Pei, Jiawei Han and etc.: Mining Access Patterns Efficiently from Web Logs, PAKDD, (2000).
6. J. Srivastava, R. Cooley, M. Deshpande and P. Tan: Web Usage Mining: Discovery and Application of Usage Patterns from Web Data, ACM SIGKDD, (2000).
7. L. Page, S.Brin, R. Motwani and T. Winograd: The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford Digital Library Technologies Project, (1998).
8. L. Wang and C. Meinel: Behaviour Recovery and Complicated Pattern Definition in Web Usage Mining. In proc. of ICWE, (2004).
9. L. Wang, C. Meinel and C. Liu: Discovering Individual Characteristic Access Patterns in Web Environment. In proc. of RSFDGrC, (2005).
10. M. Chen, J. Park: Data Mining for Path Traversal Patterns in a Web Environment. In ICDCS, (1996).
11. M. Spilioupoulou, B. Mobasher, B. Berendt and M. Nakagawa: A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis. In Journal of INFORMS on Computing, (2003).
12. Mohammed J. Zaki: Efficiently Mining Frequent Trees in a Forest. In SIGKDD (2002).
13. Perkowitz, M. and Etzioni, O. Adaptive Web Sites: Automatically Syntehsizing Web Pages. In AAAI, (1998).
14. Peter F. Brow: Class-Based n-gram Models of Natural Language. In Association for Computational Linguistics, (1992).
15. Peter Pirolli: Distributions of Surfers Path through the World Wide Web. World Wide Web 2, (1999).
16. P. Tonella: Evaluation Methods for Web Application Clustering. In proc. 9th WWW conference, (2000).
17. R. Sarukkai: Link Prediction and Path Analysis Using Markov Chains. In proc. 9th WWW conference, (2000).
18. T. Pang and V. Kurmar: Interestingness Measures for Association Patterns: A Perspercitve. Tech. Rep. University of Minnesota, (2000).
19. T. Joachims, D. Freitag and etc.: WebWatcher: A Tour Guide for the World Wide Web. In IJCAI, (1997).
20. T.R. Niesler and P.C. Woodland: Modelling Word-Pair Relations in a Category-Based Language Model. In proc. ICASSP, (1997).
21. X. Huang, F. Peng, A. An and D. Schuurmans: Dynamic Web Log Session Identification with Statistical Language Models. In Journal of American Society for Information Science and Technology, (2004).
22. Y. Fu, K. Sandhu, and M. Shih: A Generalization-based Approach to Clustering of Web Usage Sessions. In Web Usage Analysis and User Profiling, (2002).