Co-clustering for Weblogs in Semantic Space

Zong, Yu; Xu, Guandong; Dolog, Peter; Zhang, Yanchun; Liu, Renjin

Published in:
Lecture Notes in Computer Science

DOI (link to publication from Publisher):
10.1007/978-3-642-17616-6_12

Publication date:
2010

Document Version
Early version, also known as pre-print

Citation for published version (APA):
Zong, Y., Xu, G., Dolog, P., Zhang, Y., & Liu, R. (2010). Co-clustering for Weblogs in Semantic Space. *Lecture Notes in Computer Science*, *6488*, 120-127. https://doi.org/10.1007/978-3-642-17616-6_12

Co-clustering for Weblogs in Semantic Space

Yu Zong[1], Guandong Xu[2,3] ⋆, Peter Dolog[2], Yanchun Zhang[3], and Renjin Liu[1]

[1] Department of Information and Engineering, West Anhui University,China
[2] IWIS - Intelligent Web and Information Systems, Aalborg University, Computer Science Department Selma Lagerlofs Vej 300 DK-9220 Aalborg, Denmark
[3] Center for Applied Informatics, School of Engineering & Science, Victoria University, PO Box 14428, Vic 8001, Australia

Abstract. Web clustering is an approach for aggregating web objects into various groups according to underlying relationships among them. Finding co-clusters of web objects in semantic space is an interesting topic in the context of web usage mining, which is able to capture the underlying user navigational interest and content preference simultaneously. In this paper we will present a novel web co-clustering algorithm named Co-Clustering in Semantic space (COCS) to simultaneously partition web users and pages via a latent semantic analysis approach. In COCS, we first, train the latent semantic space of weblog data by using Probabilistic Latent Semantic Analysis (PLSA) model, and then, project all weblog data objects into this semantic space with probability distribution to capture the relationship among web pages and web users, at last, propose a clustering algorithm to generate the co-cluster corresponding to each semantic factor in the latent semantic space via probability inference. The proposed approach is evaluated by experiments performed on real datasets in terms of precision and recall metrics. Experimental results have demonstrated the proposed method can effectively reveal the co-aggregates of web users and pages which are closely related.

1 Introduction

Recently, the Internet becomes an important and popular platform for distributing and acquiring information and knowledge due to its rapid evolution of web technology and the influx of data sources available over the Internet in last decades [1]. Web clustering is emerging as an effective and efficient approach to organize the data circulated over the web into groups/collections in order to re-structure the data into more meaningful blocks and to facilitate information retrieval and representation [2–4]. The proposed clustering on web usage mining is mainly manipulated on one dimension/attribute of the web usage data standalone. However, in most cases, the web object clusters do often exist in the forms of co-occurrence of pages and users - the users from the same group are particularly interested in one subset of total web pages in a e-commerce site. As

⋆ corresponding author

a result, a simultaneous grouping of both sets of users and pages is more appropriate and meaningful in modeling user navigational behavior and adapting the web design, in some contexts [5–7].

Although a considerable amount of researches of co-clustering on weblogs have been done, the major approaches are developed on a basis of graph partition theory[6–8, 5]. One commonly encountered difficulty of such kind of approaches is how to interpret the cluster results and capture the underlying reason of such clustering. In contrast, latent semantic analysis is one of the effective means to capturing the latent semantic factor hidden in the co-occurrence observations. As such combining the latent semantic analysis with clustering motivates the idea presented in this paper. In this paper, we aim to propose a novel approach addressing the co-clustering of web users and pages by leveraging the latent semantic factors hidden in weblogs. With the proposed approach, we can simultaneously partition the web users and pages into different groups which are corresponding to the semantic factors derived by employing the PLSA model. In particular, we propose a new co-clustering framework, named Co-Clustering in Semantic Space (COCS) to deal with simultaneously finding web object co-clusters in a semantic space. Upon COCS, we first capture the semantic space by introducing the PLSA model; and then, we project the web objects into this semantic space via referring to the probability estimates; at last, we generate the web object co-clusters for all semantic factors in the semantic space. We conduct experimental evaluations on real world datasets.

The rest of the paper is organized as follows. In section 2, we introduce the framework of COCS. In section 3, we first introduce the PLSA model and describe the process of capturing the semantic space, and then, we discuss how to project the data objects into the captured semantic space, at last, a co-clustering clustering algorithm is proposed. To validate the proposed approach, we demonstrate experiment and comparison results conducted on two real world datasets in section 4, and conclude the paper in section 5.

2 The Framework of COCS

Prior to introduce the framework of COCS, we discuss briefly the issue with respect to sessionization process of web usage data. The user session data can be formed as web usage data represented by a session-page matrix $SP = \{a_{ij}\}$. The entry in the session-page matrix, a_{ij}, is the weight associated with the page p_j in the user session s_i, which is usually determined by the number of hit or the amount time spent on the specific page.
The framework of COCS is composed of three parts and is described as follow.

1. Learning the semantic space of web usage data;
2. Projecting the original web pages and web sessions into the semantic space respectively, i.e. each web session and page is expressed by a probability distribution over the semantic space;
3. Abstracting the co-clusters of web pages and web sessions corresponding to various semantic factors in the semantic space.

The first part is the base of COCS. For a given session-page matrix $SP = \{a_{ij}\}$, we can learn the latent semantic space Z by using the PLSA model, which is detailed in section 3.1. The main component of the second part in COCS is based on the result of first part of COCS. Two probabilistic relationship matrices of web pages and web sessions with semantic space are obtained by employing the PLSA model. Following the semantic analysis, we then project the web objects into the captured semantic space by linking these probability matrices. We further discuss this part in section 3.2.In the last part of COCS, we propose a clustering algorithm to find out the co-clusters in the semantic space by filtering out the web pages and sessions based on the probability cutting value. The detail will be described in section 3.3.

3 The Details of COCS

In this section, we will discuss the organization of each part of COCS in detail.

3.1 Capturing Semantic Space by Using PLSA

The PLSA model is based on a statistic model called aspect model, which can be utilized to identify the hidden semantic relationships among general co-occurrence activities. Similarly, we can conceptually view the user sessions over web page space as co-occurrence activities in the context of web usage mining to discover the latent usage pattern. For the given aspect model, suppose that there is a latent factor space $Z = \{z_1, z_2, \cdots z_K\}$ and each co-occurrence observation data (s_i, p_j) is associated with the factor $z_k \in Z$ by a varying degree to z_k.

From the viewpoint of aspect model, thus, it can be inferred that there are existing different relationships among web users or pages related to different factors, and the factors can be considered to represent the user access patterns. In this manner, each observation data (s_i, p_j) can convey the user navigational interests over the K-dimensional latent factor space. The degrees to which such relationships are explained by each factor derived from the factor-conditional probabilities. Our goal is to discover the underlying factors and characterize associated factor-conditional probabilities accordingly.

By combining probability definition and Bayesian formula, we can model the probability of an observation data (s_i, p_j) by adopting the latent factor variable z as:

$$P(s_i, p_j) = \sum_{z_k \in Z} P(z_k) \cdot P(s_i|z_k) \cdot P(P_j|z_k) \tag{1}$$

Furthermore, the total likelihood of the observation is determined as

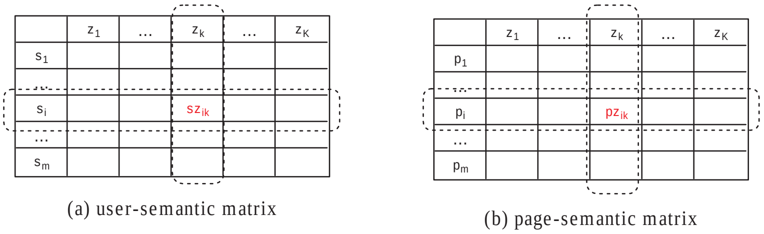
$$L_i = \sum_{s_i \in S, p_j \in P} m(s_i, p_j) \cdot \log P(s_i, p_j) \tag{2}$$

where $m(s_i, p_j)$ is the element of the session-page matrix corresponding to session s_i and page p_j.

We utilize Expectation Maximization (EM) algorithm to perform the maximum likelihood estimation in latent variable model [9], i.e. the probability distributions of $P\left(s_i|z_k\right)$, $P\left(p_j|z_k\right)$ and $P(z_k)$.

3.2 Projecting Data Objects into Semantic Space

For each user session s_i, we can compute a set of probabilities $P\left(z_k\left|s_i\right.\right)$ corresponding to different semantic factors via the Bayesian formula.
With the learned latent semantic factor space, it is noted that for any semantic factor $z_k \in Z$, each user session has a probability distribution with it. And the web page has the similar relationship with each semantic factor as well. For brevity we use matrix SZ to represent the relationship between all the user sessions and the semantic space Z, and PZ denotes the relationship between all the pages and the semantic space Z. Fig.1 shows these relationships in schematic structure of matrix. Fig.1(a) represents the matrix SZ and Fig.1(b) shows the matrix PZ respectively. In each matrix, the element represents the probabilistic relatedness of each session or page with each semantic factor. In summary after the semantic space learning, the original user session and web page are simultaneously projected onto a same semantic factor space with various probabilistic weights.



(a) user-semantic matrix

(b) page-semantic matrix

Fig. 1. relationship between user, page and semantic factor

Since we aim to find out the co-cluster of pages and sessions corresponding to each semantic factor, so we join the matrix SZ and PZ together into a single $SPZ = \begin{bmatrix} SZ \\ PZ \end{bmatrix}$. Then, we apply a co-clustering algorithm, which will be presented in section 3.3 to find out the co-clusters in the semantic space by referring to the calculated matrices .

3.3 Abstracting Co-clustering in Semantic Space

In this section, a clustering algorithm is proposed to find out co-clusters embedded in semantic space. For each semantic $z_k \in Z$, there is a corresponding co-cluster C_k. Since web pages and web sessions have a latent semantic relationship with each semantic factor in probability model, so we use Definition 1 to define a co-cluster C_k.

Definition 1. *Given a semantic $z_k \in Z$, and a probability threshold μ, the co-cluster C_k corresponding to the semantic factor z_k is defined as: $C_k = (S_k, P_k)$*

where $S_k = \{s_i | sz_{ik} \geq \mu, i = 1, \ldots, m\}$ and $P_k = \{p_j | pz_{jk} \geq \mu, j = 1, \ldots, n\}$. We use hard partition method to generate co-clusters in semantic space. In the semantic space Z, we presume there are K co-clusters with respect to K semantic factors. Definition 2 shows the process of obtaining the co-clustering results in the semantic space Z.

Definition 2. *Given a co-cluster C_k corresponding to a semantic factor $z_k \in Z$, $k = 1, \ldots, K$. The co-clustering result C in the semantic space Z is defined as: $C = \{C_1, C_2, ..., C_K\} = \{(S_1, P_1), \cdots, (S_k, P_k)\}$*

Algorithm 1 shows the details of finding co-clusters. At first, we set $C \leftarrow \emptyset$ as the initialization after the semantic factor space is learned via the PLSA model, and then, for each semantic factor z_k, we simultaneously filter out the web sessions and web pages which have the occurrence probability to the factor z_k greater than a probability threshold μ, and the co-cluster C_k corresponding to z_k is generated by aggregating these filtered sessions and pages. At last, we insert C_k into C.

Algorithm 1. Discovering co-clusters in the semantic space.

```
Input: the web usage matrix $SP$, the probability threshold $\mu$
Output: $C$
1: employ the PLSA model to capture the latent semantic space $Z$, and obtain the projection matrices of web sessions and pages $SZ$ and $PZ$ over the semantic space and join them together;
2: $C \leftarrow \emptyset$;
3: for $k = 1, \ldots, K$
    3.1 find out the co-cluster $C_k$ corresponding to the semantic factor $Z_k$ according to Definition 1;
    3.2 $C \leftarrow C \cup C_k$;
4: end
5: return $C = \{C_1, C_2, ..., C_K\}$.
```

For each semantic z_k in Z, we must check the probability matrix SZ and PZ to decide whether their probabilities are over the probability threshold μ. So it needs $O(m + n)$ time. Algorithm 1 needs to check the K semantic factors in Z, so the time cost is $O(K(m + n))$.

4 Experiments and Evaluations

In order to evaluate the effectiveness of our proposed method, we have conducted preliminary experiments on two real world data sets. The first data set we used is downloaded from KDDCUP website [4]. After data preparation, we have setup

[4] http://www.ecn.purdue.edu/KDDCUP/

an evaluation data set including 9308 user sessions and 69 pages and we refer this data set to "KDDCUP data". The second data set is from a university website log file [10]. The data is based on a 2-week weblog file during April of 2002 and the filtered data contains 13745 sessions and 683 pages. For brevity we refer this data as "CTI data". By considering the number of web pages and the content of the web site carefully and referring the selection criteria of factors in [11, 10], we choose 15 and 20 factors for KDDCUP and CTI dataset for experiment, respectively.

Our aim is to validate how strong the correlation between the user sessions and pages within the co-cluster is. Here we assume a better co-clustering representing the fact that most of the user sessions visited the pages which are from the same co-cluster whereas the pages were largely clicked by the user sessions within the same subset of sessions and pages. In particular we use precision, recall and F-score measures to show how likely the users with similar visiting preference are grouped together with the related web pages within the same cluster. For each $C_k = (S_k, P_k)$, its precision and recall measures are defined as below.

Definition 3. *Give a co-cluster $C_k = (S_k, P_k)$, its precision and recall measures are defined as the linear combination of the row (i.e.session) precision and column (i.e. page) precision, and the linear combination of the row and column recall.*
$precision(C_k) = \alpha * precision(R_{C_k}) + (1 - \alpha) * precision(C_{C_k})$
$recall(C_k) = \alpha * recall(R_{C_k}) + (1 - \alpha) * recall(C_{C_k})$

where the row and column precision, and the row and column recall are defined as follows, respectively
$precision(R_{C_k}) = \frac{\sum_{i \in S_k} \sum_{j \in P_k} a'_{ij}/|P_k|}{|S_k|}$, $precision(C_{C_k}) = \frac{\sum_{j \in P_k} \sum_{i \in S_k} a'_{ji}/|S_k|}{|P_k|}$,
$recall(R_{C_k}) = \sum_{i \in S_k} \frac{\sum_{j \in P_k} a'_{ij}}{\sum_{l=1}^{n} a'_{il}}/|S_k|$, $recall(C_{C_k}) = \sum_{j \in P_k} \frac{\sum_{i \in S_k} a'_{ji}}{\sum_{l=1}^{n} a'_{jl}}/|P_k|$,
where α is the combination factor, a'_{ji} is the binary representation of usage data, that is, $a'_{ji} = 1$, if $a_{ij} \geq 1$; is 0, otherwise.

Definition 4. *Given the co-clustering of $C = \{C_1, ..., C_K\}$, the precision and recall of C is defined as follows:*
$precision = \frac{1}{K} \sum_{k=1}^{K} precision(C_k)$, $recall = \frac{1}{K} \sum_{k=1}^{K} recall(C_k)$

Because we consider that sessions and pages have the equal contribution on the cluster conformation, in the experiment, the combination factor α is set as 0.5. It is clear that, the higher value of precision and recall denotes a better clustering being executed. We conducted evaluations on these two datasets in terms of precision, recall. We run 30 times of the proposed algorithm COCS and the compared algorithm: Spectral Co-Clustering (SCC) on CTI and KDDCUP datasets first, and then we denote the process as COCS_CTI, COCS_KDD, SCC_CTI and SCC_KDD, respectively. The results are presented in Fig.2-3.

From Fig.2, it is seen that the precisions of COCS on CTI and KDDCUP data-sets are between 0.82 to 0.86 while the precisions of SCC on these two datasets are ranging from 0.74 to 0.81, resulting in by 7.7% worse than COCS. In COCS, sessions and pages are organized according to different semantic factors, that is, each semantic factor in the semantic space has a corresponding co-cluster. However, in SSC, the co-clusters are generated by calculating the cosine similarity between sessions and pages on a projected spectrum (i.e attribute) space. The co-cluster of sessions and pages in same co-cluster have no straight relationship with the semantic factor. According to the co-clustering result of COCS and SCC, we conclude the fact that co-clusters of COCS are grouped more semantic-related than those of SSC, so the precision values of COCS on these two datasets are both higher than that of SCC.

Fig.3 shows the recall comparison of COCS and SCC on CTI and KDDCUP datasets. From this figure, we can know that, in addition to the finding that COCS always outperforms SSC by around 15%, the recall values of COCS and SCC on the CTI dataset are much higher than those on the KDDCUP dataset respectively - recall values ranging in (0.58-0.62) vs. (0.49-0.55). This is likely because that the KDD dataset is relative sparse due to the bigger granularity level.

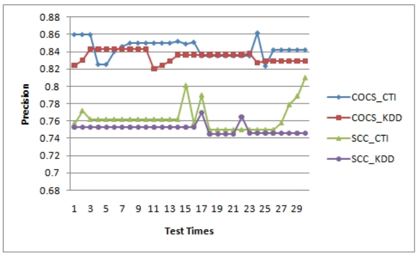


Fig. 2. Precision Comparison Results

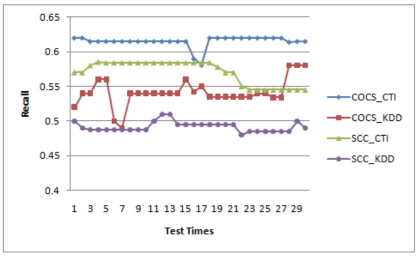


Fig. 3. Recall Comparison Results

5 Conclusion and Future work

Web clustering is an approach for aggregating web objects into various categories according to underlying relationships among them. In this paper, we address the question of discovering the co-clusters of web users and web pages via latent semantic analysis. A framework named COCS is proposed to deal with the simultaneous grouping of web objects. Experiments results have shown that the proposed method largely outweighs the existing co-clustering approaches in terms of precision and recall measures due to the capability of latent semantic analysis. Our future work will focus on the following issues: we intend to conduct more experiments to validate the scalability of our approach, and investigate the impact of the selection of different semantic factor number on the co-clustering performance.

6 Acknowledgment

This work has been partially supported by FP7 ICT project "M-Eco: Medical Ecosystem Personalized Event-Based Surveillance" under grant No. 247829 and "KiWi - Knowledge in a Wiki" under grant agreement No. 211932, and the Nature Science Foundation of Anhui Education Department under the Grant No KJ2009A54, KJ2010A325 and KJ2009B126.

References

1. Zhang, Y., Yu, J.X., Hou, J.: Web Communities: Analysis and Construction. Springer, Berlin Heidelberg (2006)
2. Wang, X., Zhai, C.: Learn from web search logs to organize search results. In: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2007) 87–94
3. Kummamuru, K., Lotlikar, R., Roy, S., Singal, K., Krishnapuram, R.: A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In: WWW '04: Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, ACM (2004) 658–665
4. Flesca, S., Greco, S., Tagarelli, A., Zumpano, E.: Mining user preferences, page content and usage to personalize website navigation. World Wide Web Journal **8**(3) (2005) 317–345
5. Zeng, H.J., Chen, Z., Ma, W.Y.: A unified framework for clustering heterogeneous web objects. In: WISE '02: Proceedings of the 3rd International Conference on Web Information Systems Engineering, Washington, DC, USA, IEEE Computer Society (2002) 161–172
6. Xu, G., Zong, Y., Dolog, P., Zhang, Y.: Co-clustering analysis of weblogs using bipartite spectral projection approach. In: Proceedings of 14th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. (2010)
7. Koutsonikola, V.A., Vakali, A.: A fuzzy bi-clustering approach to correlate web users and pages. I. J. Knowledge and Web Intelligence **1**(1/2) (2009) 3–23
8. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review **69** (2004) 026113
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B **39**(1) (1977) 1–38
10. Jin, X., Zhou, Y., Mobasher, B.: A maximum entropy web recommendation system: Combining collaborative and content features. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'05), Chicago (2005) 612–617
11. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning Journal **42**(1) (2001) 177–196