Using Subspace Analysis for Event Detection from Web Click-through Data

Ling Chen
L3S Research Center
University of Hannover
30167 Hannover, Germany
lchen@l3s.de

Yiqun Hu
School of Computer Engg
Nanyang Technological University
Singapore 639798
y030070@ntu.edu.sg

Wolfgang Nejdl
L3S Research Center
University of Hannover
30167 Hannover, Germany
nejdl@l3s.de

ABSTRACT

Although most of existing research usually detects events by analyzing the content or structural information of Web documents, a recent direction is to study the usage data. In this paper, we focus on detecting events from Web *click-through data* generated by Web search engines. We propose a novel approach which effectively detects events from click-through data based on robust subspace analysis. We first transform click-through data to the $2D$ polar space. Next, an algorithm based on Generalized Principal Component Analysis (GPCA) is used to estimate subspaces of transformed data such that each subspace contains query sessions of similar topics. Then, we prune uninteresting subspaces which do not contain query sessions corresponding to real events by considering both the semantic certainty and the temporal certainty of query sessions in each subspace. Finally, various events are detected from interesting subspaces by utilizing a nonparametric clustering technique. Compared with existing approaches, our experimental results based on real-life click-through data have shown that the proposed approach is more accurate in detecting real events and more effective in determining the number of events.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

General Terms

Algorithms, Measurement, Experimentation

Keywords

click-through data, event detection, subspace estimation, GPCA

1. INTRODUCTION

With the prevailing publishing activities over the Internet, the Web of nowadays covers almost every object and event in the real world. This phenomenon has recently motivated the event detection community to discover knowledge such as topics, events and stories from large volumes of Web data [4][3]. Most of existing work detect events from either the content [4] or the structures [3] of Web data. A recent research direction is to study the Web usage data. For example, Zhao et al. [7] proposed to detect events from Web click-through data, which are the log data generated by Web search engines. In this paper, we also focus on detecting events from Web click-through data.

2. EVENT DETECTION

Each entry of Web click-through data basically records the following four types of information: an anonymous user identity, the query issued by the user, the time at which the query was submitted for search, and the URL of clicked search result [5]. Hence, Web click-through data at least provide two aspects of useful knowledge for event detection: *semantics of events* (e.g., the knowledge indicated by the queries and the corresponding clicked pages) and *time of events* (e.g., the knowledge indicated by the timestamps at which the queries are issued). Given a collection of Web click-through data, the overview of our approach is presented in Figure 1. Basically, there are four steps involved: *polar transformation*, *subspace estimation*, *subspace pruning*, and *cluster generation*.

1. **Polar transformation**. Firstly, we transform click-through data to 2D polar space. Each *query session*, containing a query and a set of corresponding pages clicked by a user, is mapped to a point in polar space such that the angle θ and radius r of the point respectively reflect the semantics and the occurring time of the query session. Particularly, given a set of query sessions $\{S_1, S_2, \cdots, S_n\}$, the radius of the point (θ_i, r_i) corresponding to the query session S_i is given by

$$r_i = \frac{T(S_i) - \min_j(T(S_j))}{\max_j(T(S_j)) - \min_j(T(S_j))}$$

where $T(S_i)$ is the occurring time of query sessions S_j. r_i takes value in the range of $[0, 1]$.

We define the semantic similarity between two query sessions $S_1 = (Q_1, P_1)$ and $S_2 = (Q_2, P_2)$ as

$$Sim(S_1, S_2) = \frac{\alpha \times |Q_1 \cap Q_2|}{max\{|Q_1|, |Q_2|\}} + \frac{(1-\alpha) \times |P_1 \cap P_2|}{max\{|P_1|, |P_2|\}}$$

where Q_i and P_i are a set of query keywords and a set of clicked pages of the session S_i respectively. We then compute a semantic similarity matrix for the set of query sessions and perform PCA on the matrix. We use the first principle component to preserve the dominant variance in semantic similarities. Let $\{f_1, f_2, \cdots, f_n\}$ be the first principal component which corresponds to the set of query sessions $\{S_1, S_2, \cdots, S_n\}$. A query session S_i can be mapped to a point (θ_i, r_i) where θ_i is computed as

$$\theta_i = \frac{f_i - \min_j(f_j)}{\max_j(f_j) - \min_j(f_j)} \times \frac{\pi}{2}$$

Obviously, θ_i is restricted to $[0, \pi/2]$.

2. **Subspace estimation**. Based on our polar transformation, query sessions of similar semantics should be mapped to points of similar angles and lie on one and only one 1D subspace. Therefore, in this step, we perform subspace estimation on the set of transformed data. Our algorithm is based on Generalized Principal Component Analysis (GPCA) [6], an algebro-geometric approach

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

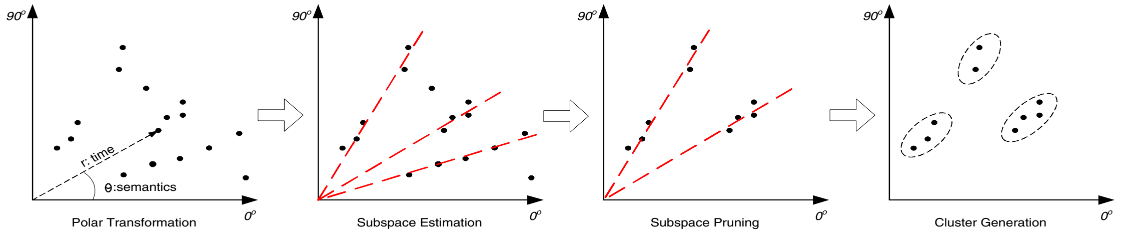

Figure 1: Overview of DECK.

which simultaneously estimates subspace bases and assigns data points to subspaces. We noticed that the performance of GPCA degrades in the presence of outliers and noises. We then improve robustness of GPCA by assigning weight coefficients to data points and demoting the impact of noises and outliers. As a true data point should locate in a cluster inside a subspace, its K nearest neighbors should have small variance in both the subspace direction and the orthogonal direction of the subspace. Hence, given a data point x_i, we assign a weight $W(x_i)$ as follows,

$$W(x_i) = \frac{1}{1 + (s(NN_{x_i}) + n(NN_{x_i})) \times \frac{s(NN_{x_i})}{n(NN_{x_i})}} \qquad (1)$$

where $s(NN_{x_i})$ is the variance of x_i's K nearest neighbors along the subspace direction and $n(NN_{x_i})$ is the variance of its neighbors along the orthogonal direction of the subspace. When the data point x_i lies in a cluster where data points spread along the direction of the subspace, both the value of $\frac{s(NN_{x_i})}{n(NN_{x_i})}$ and the value of $s(NN_{x_i}) + n(NN_{x_i})$ are small. Hence, the weight $W(x_i)$ is close to 1. Otherwise, $\frac{s(NN_{x_i})}{n(NN_{x_i})}$ and/or $s(NN_{x_i}) + n(NN_{x_i})$ are large, which results in a small $W(x_i)$.

3. **Subspace pruning**. Not every subspace is interesting such that it contains clusters corresponding to real events. Hence, we prune uninteresting subspaces in this step. Based on our polar transformation schemes, the temporal “burst" and the semantic “burst" of query sessions should be reflected by the *certainly* distribution of data points along the subspace direction and the orthogonal direction of the subspace respectively. In order to measure the certainty of the distribution of data points along the two directions, we project data points to the two directions respectively and calculate the respective histograms of the distributions. Let $\langle h_1, h_2, \cdots, h_m\rangle$ and $\langle v_1, v_2, \cdots, v_n\rangle$, where h_i and v_i are individual bins, be the two corresponding histograms. We employ the *entropy* measure to define the interestingness of a subspace s_i as follows.

$$I(s_i) = 1 - [-p\sum_{i=1}^{m} h_i \log h_i - (1-p)\sum_{i=1}^{n} v_i \log v_i] \qquad (2)$$

where $p \in [0, 1]$ is a weight which adjusts the importance of the entropy values in the two directions. The interestingness measure takes values from 0 to 1. The more certain the distributions in two directions, the smaller the entropies in the brackets of equation (2), the greater the value of interestingness. Given some threshold ζ, subspace s_i will be pruned if $I(s_i) < \zeta$.

4. **Cluster generation**. After pruning uninteresting subspaces, events can be detected from the remaining subspaces by clustering. Particularly, we detect various events from interesting subspaces by employing a non-parametric clustering method called Mean Shift [2].

3. EXPERIMENTS & CONCLUSIONS

We conduct experiments on the real-life Web click-through data collected by AOL [5] from March 2006 through May 2006. We manually labelled a set of events from the data set. After filtering events which are represented by less than 50 query sessions, a total of 35 events are used in our experiments. The complete list of events is given in [1]. We then randomly select query sessions which do not represent any real events, together with the query sessions corresponding to real events, to generate five data sets, which respectively contain 5K, 10K, 20K, 50K and 100K query sessions.

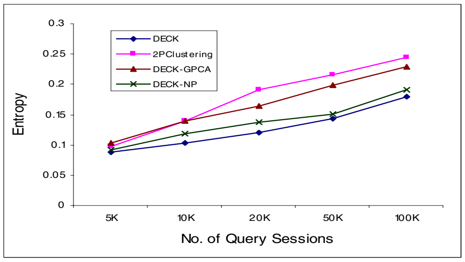

Figure 2: Entropy comparison between algorithms.

One of our performance evaluation using entropy measure is shown in Figure 2. For each generated cluster i, we compute p_{ij} as the fraction of query sessions (or query-page pairs for the existing approach [7]) representing the true event j. Then, the entropy the of cluster i is $E_i = -\sum_j p_{ij} \log p_{ij}$. The total entropy can be calculated as the sum of the entropies of each cluster weighted by the size of each cluster: $E = \sum_i^m \frac{n_i \times E_i}{n}$, where m is the number of clusters, n is total number of query sessions (query-page pairs) and n_i is the size of cluster i. As shown by the figure, our approach (denoted as DECK in the figure) works better than the existing approach (denoted as 2PClustering in the figure). The figure also reveals that our approach outperforms two of its alternative versions: DECK-GPCA (which does not improve the robustness of GPCA) and DECK-NP (which does not prune uninteresting subspaces).

In general, we proposed a novel approach for detecting events from Web click-through data. Our approach based on robust subspace analysis considers the temporal feature and semantic feature of query sessions simultaneously. Experiments on real-life Web click-through data [5] showed the effectiveness of the proposed approach.

4. REFERENCES

[1] Technical report. In *http://www.l3s.de/˜ lchen/TR/deck.pdf*.

[2] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. In *IEEE TPAMI*, volume 24, 2002.

[3] W.-S. Li, K. S. Candan, Q. Vu, and D. Agrawal. Retrieving and organizing Web pages by “information unit”. In *WWW*, 2001.

[4] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In *SIGIR*, 2005.

[5] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *The First International Conference on Scalable Information Systems*, 2006.

[6] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis. In *IEEE CVPR*, 2003.

[7] Q. Zhao, T.-Y. Liu, S. S. Bhowmick, and W.-Y. Ma. Event detection from evolution of click-through data. In *KDD*, 2006.