Clustering for Web Information Hierarchy Mining

Hung-Yu Kao, Jan-Ming Ho*, and Ming-Syan Chen

Electrical Engineering Department
National Taiwan University
Taipei, Taiwan, ROC
E-Mail: {bobby@arbor.ee.ntu.edu.tw, mschen@cc.ee.ntu.edu.tw}

*Institute of Information Science**
Academia Sinica
Taipei, Taiwan, ROC
E-Mail: hoho@iis.sinica.edu.tw

Abstract

Benefiting from the growth of techniques of dynamic page generation, the amount and the complexity of Web pages increase explosively. The structures of Web pages which are dynamically generated by the same templates are thus similar to one another and are usually assembled by a set of fundamental information clusters These neighboring information clusters usually represent the similar semantics and form a larger cluster with the more generalized information. The hierarchical structure generated by information clusters in a bottom-up manner is called the information hierarchy of a page. In this paper, we study the problem of mining the information hierarchies of pages in Web sites to recognize the information distribution of pages within the multi-level, multi-granularity configurations. Explicitly, we propose an information clustering system that applies a top-down ***information centroid*** searching algorithm and a multi-granularity centroid converging process on the document object model (***DOM***) trees of pages to build the information hierarchies of pages. Experiments on several real news Web sites show the high precision and recall rates of the proposed method on determining information clusters of pages and also validate its practical applicability to real Web sites.

1. Introduction

Benefiting from the growth of techniques of dynamic page generation, the amount and the complexity of Web pages increase explosively. Many Web pages are online generated for the purposes of maintenance, flexibility, and scalability of Web sites. These Web sites are referred to as *systematic* Web sites in [3]. The structures of most pages in systematic Web sites are dynamically generated by the same templates. These structures are therefore similar to one another and are usually assembled by a set of fundamental information clusters. An ***information cluster*** is defined as a sub-structure of a page which provides a unique semantic representation to users among pages in a Web site and is composed of information elements or smaller information clusters, where an ***information element*** is one context or an anchor with a non-zero length. In this paper, one information element providing good information for users is called an ***information authority***. In contrast, an information element is called an ***information hub*** if it contains information linking to information authorities. The definitions of the information authority and the information hub are similar to those of the hub and the authority in [4], but different from them in that information authorities / hubs here are not specific to any topic.

We define the information scale of an information element as the amount of information provided to users. Note that in a Web site, the information scales and characteristics of the neighboring information clusters are usually similar to one another. In view of this, we can merge them into a larger block to represent more generalized information. Such a merged structure is considered as the high-level information cluster corresponding to these merged clusters. Such merging is referred to as ***information clustering***. After the clustering, a page composes of several disjoining information clusters. We define the configuration formed by the set of information clusters after the k-th level information clustering as the k-th level clustering, denoted by L_k. The information hierarchy is then built by Clusterings L_0, L_1, …, L_n, where Clustering L_0 corresponds to the configuration of sets of all information elements, whereas Clustering L_n is the configuration for the converged clustering, i.e., $L_n = L_{n+1} = L_{n+2}$ and so forth.

In an HTML document, tags are inserted for purposes of the page layout, content presentation and for providing interactive functions. In this paper, we extract and utilize the knowledge in the tagging tree structure, or referred to the **D**ocument **O**bject **M**odel [8], i.e., **DOM**, of a Web page and apply the information theory to mine the information hierarchy. Specifically, we propose in the paper an information clustering system which builds the information hierarchy of each page in a Web site according to both the knowledge of the information contained and the structure of pages automatically.

The main mining flow in the proposed system is to use the information theory to evaluate the information amount of content and sub-structures, and then to construct the information hierarchy by applying the specific clustering methods. We first scatter the original DOM tree into several small and non-overlapped sub-trees. The system then applies a top-down searching algorithm to select a set of top-n sub-trees referred to as the information centroids in the paper. Information clusters with different levels are built by converging the centroids using the proposed bottom-up multi-granularity centroid converging (*MGCC*)

Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03)
0-7695-1932-6/03 $17.00 © 2003 **IEEE**

process. The information hierarchy is then constructed by the set of all configurations which contain information clusters with different levels. Experiments on several real news Web sites show the high precision and recall rates of the proposed method *MGCC* and also validate its practical applicability to real Web sites.

The remainder of this paper is organized as follows. In Section 2, we describe the related works. The proposed system is described in Section 3. In Section 4, we empirically evaluate the performance of the proposed system by several real news Web sites. The paper concludes with Section 5.

2. Related works

The research in [5] provides a mechanism to construct the multi-granularity and topic-focused Web site maps. The constructed site map can be considered as a site-level information hierarchy, different from the proposed page-level information hierarchy in the paper.

Works in [1][2] also provide auxiliary systems to help the information extraction of semistructure documents. However, they need either the pre-marked training set or a considerable amount of human labor involved to process the information extraction semi-automatically. It is noted that there are also works on mining informative structure [3][6], which are, however, different from our work in that these prior works mainly dealt with the mining blocks delimited by <TABLE> tags. In contrast, we would like to mine the fine-grained blocks using the DOM tree.

3. The information clustering on the DOM

We develop a DOM-based information clustering system to build the information hierarchy of each page in a Web site according to the knowledge in the tree structures of pages automatically. The mining flow of the system, shown in Figure 1, consists of two main phases, i.e., (1) the information coverage tree building phase, and (2) the multi-granularity information clustering phase. We will describe these two phases with more details in Section 3.1 and Section 3.2 respectively.

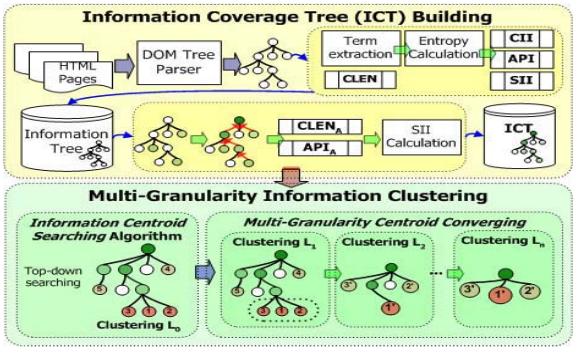

Figure 1: The system flow of information clustering

3.1 Phase 1: Information Coverage Tree Building

According to the *innerText*, i.e., the context delimited by the tag, and structure of a page, we utilize some features of nodes to indicate the scales of the information authority and the information hub, namely (1) the content length (***CLEN***), (2) the content information index (***CII***), (3) the anchor precision index (***API***), and (4) the structure information index (***SII***). We then define a tree with the bottom-up aggregated features as an information coverage tree (abbreviated as ***ICT***). In the proposed system, we aggregate node information *CLEN* and *API* to get the corresponding aggregated features, denoted by $CLEN_A$ and API_A. We use these two normalized aggregated values to indicate the aggregated scales of information authority and information hub of a node.

For the calculations of these features, we parse the *innerText* of the root node to extract meaningful terms. A term corresponds to a meaningful keyword or phrase. After extracting terms in all crawled pages, we calculate the entropy value of each term according to its term frequency. From Shannon's information entropy [7], the entropy of term $term_i$ can be formulated as:

$$EN(term_i) = -\sum_{j=1}^{n} w_{ij} \log_n w_{ij}, \text{ where } n = |D|, D \text{ is the set of pages},$$

in which w_{ij} is the value of normalized term frequency in the page set. When entropy values of terms are calculated, we average the entropy values of terms in an *innerText* of node N to get *CII(N)*, the content information index of N, i.e.,

$$CII(N) = \frac{\sum_{j=1}^{k} EN(term_j)}{k}, \text{ where } \forall_{j=1\sim k} term_j \text{ in innerText of } N.$$

The *CII* value of node N represents the amount of information carried in a sub-tree rooted by N. We also define the value of the anchor precision index to indicate the correlation of the anchor and its linking page. We use the anchor text to evaluate the value of *API*. The correlation index *API* is defined as, $API(N) = \sum_{j=1}^{m} \frac{1}{EN(term_j)}$, where $term_j$ is the term concurrently appearing in both the anchor text of N and the linked page and m is the number of matched terms.

Finally, the index *SII* of a node is calculated according to the distribution of the feature values of the node's children. We define *children(N)* as the set of all non-dummy children of the node A. We define the SII value of node N with children $n_0, n_1, \ldots, n_{m-1}$ for feature f_i as:

$$SII(N, f_i) = -\sum_{j=0}^{m-1} w_{ij} \log_m w_{ij}, \text{ where } w_{ij} = \frac{f_i(n_j)}{\sum_{k=0\sim m-1} f_i(n_k)}, \forall n_k \in children(N).$$

We apply entropy calculation here to represent the distribution of children's feature values of any node with

Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03)
0-7695-1932-6/03 $17.00 © 2003 **IEEE**

more than one child. The value of *SII* indicates the degree that the feature values of the node are dispersed among its children. When the value of $SII(N, f_i)$ is higher, the values of all children's f_i tend to be equal.

3.2 Phase 2: Multi-Granularity Information Clustering

In this phase, we first apply the proposed information centroid searching algorithm to select the top-k information clusters of a page, which are called information centroids. The bottom-up centroid converging method is then applied to these top-k centroids to build the multi-granularity information clusterings in different levels. After scattering a DOM tree into a set of sub-trees by the given SII constraint, i.e., **S**II **T**hreshold (***ST***), we can find the set of scattered centroid candidates. According to the covering characteristic of the aggregated features, a top-down, greedy searching algorithm can extract information centroids with the top-k information scales among these candidates.

After the top-k information centroids are extracted, we then apply a node merging process, called centroid converging, to merge the neighboring and similar information centroids and clusters into a more generalized cluster. One iteration of the basic centroid converging process in the i-th level converging contains two steps, including (1) verifying the incremental cluster constraint, i.e., $\boldsymbol{CInfo_{base}} + \mathbf{i} * \boldsymbol{CInfo_{inc}}$ and (2) finding the new converged centroid. The values in the tuple ($\boldsymbol{CInfo_{base}}$, $\boldsymbol{CInfo_{inc}}$) are pre-assigned thresholds and are used for the judgments of clustering a set of information centroids and continuing to converge in a level of converging. Note that a level of converging process may contain more than one iteration of the basic process as shown in Figure 2.

We define a converging scope (or abbreviatedly scope) as a set of sibling nodes in the DOM tree, which contains at least one centroid. Each node *N* in the converging scopes of the *MGCC* process can be expressed as a node in the 2-dimensional space of the scales of information authority and information hub with the tuple value $(\frac{API_A(N)}{AC(N)}, CII(N))$ where *AC(N)* is the count of anchors contained in *T(N)*. For a set of centroids in the scope k, we calculate the value $\boldsymbol{CInfo_k}$, which is equal to the geometric average of the maximum difference of information authorities, i.e., $\boldsymbol{CInfoAuth_k}$, and the maximum differences of information hubs, i.e., $\boldsymbol{CInfoHub_k}$, among the set of centroids and the converged centroid in the scope as shown in Figure 3. We use the value $\boldsymbol{CInfo_k}$ to measure the information diversity between centroids in the same scope. $\boldsymbol{CInfo_k}$ is equal to the distance between the centroid and the converged centroid in the 2-dimensional space when there is only one centroid in the scope.

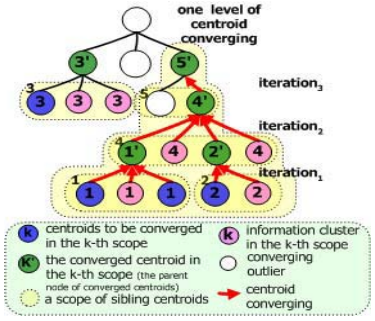

Figure 2: the centroid converging in the DOM tree

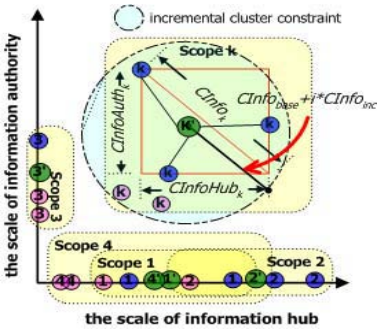

Figure 3: The different converging cases of the scopes in Figure 2

4. Clustering results and evaluations

News Web sites are typical systematic Web sites. The structures of TOCs and articles pages are distinct and are very appropriate to evaluate the proposed method of mining the information hierarchy. We therefore conduct our experiments on pages in the datasets used in [3]. The datasets contain several commercial news Web sites as described in Table 1.

Table 1: Datasets for experiments and evaluations of information clusters

Site Abbr.	URL	Total pages	TOC pages	TOC answer	article answer
CDN	www.cdn.com.tw	261	25	22*/38	60#/63
TIMES	news.chinatimes.com	3747	79	69/313	66/68
CNA	www.cna.com.tw	1400	33	29/106	50/50
CNET	taiwan.cnet.com	4331	78	38/84	37/86
CTS	www.cts.com.tw	1316	31	19/21	53/80
TVBS	www.tvbs.com.tw	740	13	12/25	50/50
TTV	www.ttv.com.tw	861	22	18/20	42/75
UDN	udnnews.com	4676	252	243/674	52/106
TOTAL		12035	530	450/1281	411/579

#: The domain experts selected the article pages with different and distinctive tagging styles to be the article answer set.

4.1 The results of information clustering

We apply the proposed information clustering method on the pages with the marked types to build their information hierarchies. After building the *ICT*s, we apply the information centroid searching algorithm under different

Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03)
0-7695-1932-6/03 $17.00 © 2003 **IEEE**

SII thresholds. We use the different *ST* values to control the number and granularity of the information centroids as shown in Figure 4.

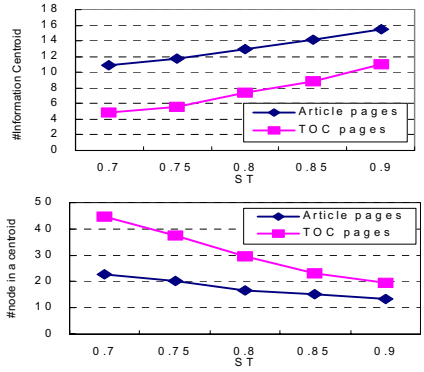

Figure 4: the distribution of numbers of information centroid and corresponding sizes under different STs.

4.2 Evaluations on informative information clusters

To assess the proposed process, we use two answer sets, i.e., TOC blocks and article blocks, to evaluate the precision and recall rates of Clustering L_1. Figure 5 shows that the average improvement of the 1-st level converging. We use two evaluation methods, i.e., significant node coverage (SNC) and information coverage (IC) to evaluate the precision and recall rates of TOC and article pages respectively. Explicitly, SNC evaluates the precision (P) and recall (R) rates by matching anchor nodes and IC matches *innerText*. We also use the F-measure which is the harmonic mean of values of precision and recall and is formulated as $\frac{2*(R*P)}{R+P}$ to evaluate results in a single efficiency measure. It can be observed from this figure that the enhanced performance of the information hierarchy is prominent when k=1.

5. Conclusion

In this paper, we propose an information hierarchy mining system that applies the multi-granularity centroid converging information clustering on the DOM trees of pages. With a DOM tree scattered into many small pieces of sub-trees by a dynamic threshold of the structure information after the aggregated features are computed, the system applies a top-down information centroid searching algorithm to select a set of sub-structures. The information hierarchy is then built by the multi-level configurations which are generated from expanding and merging the centroids using the proposed multi-granularity centroid converging method (***MGCC***). The attained information hierarchy is not only useful for search engines, inter-media information agents, and crawlers to index, extract and navigate significant information from a Web site, but also for providing the hierarchical configurations of a page according to the amount of information contained. The clustering results show that the proposed process can effectively extract the information hierarchies of pages and experiments on several real news Web sites show the high precision and recall rates of the proposed system on finding information clusters of pages and also validate its practical applicability to real Web sites.

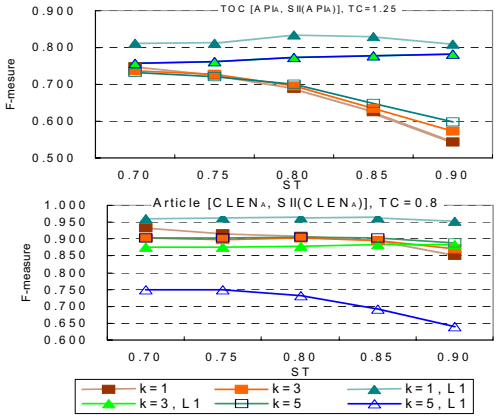

Figure 5: The evaluation of the informative clusters in Clustering L_1

Acknowledgement

The authors are supported in part by the Ministry of Education Project No.89-E-FA06-2-4, and the National Science Council Project No. NSC 91-2213-E-002-034 and NSC 91-2213-E-002-045, Taiwan, Republic of China.

References

[1] B. Adelberg. NoDoSE—a tool for semi-automatically extracting structured and semistructured data from text documents. Proc. of the 1998 ACM SIGMOD international conference on Management of data (SIGMOD'98), 1998.

[2] C. N. Hsu and M. T. Dung. Generating Finite-state Transducers for Semi-structured Data Extraction from the Web. Information Systems, 23(8):521-538, 1998.

[3] H.-Y. Kao, S.-H. Lin, J.-M. Ho and M.-S. Chen. Entropy-Based Link Analysis for Mining Web Informative Structures. Proc. of the ACM 11th International Conf. on Information and Knowledge Management (CIKM-02), Nov. 4-9, 2002.

[4] J. M. Kleinberg, Authoritative sources in a hyperlinked environment. ACM-SIAM Symposium on Discrete Algorithms. 1998.

[5] W. S. Li, N. F. Ayan, O. Kolak and Q. Vu, Constructing Multi-Granular and Topic-Focused Web Site Maps, Proc. of the 10th World Wide Web Conference, 2001.

[6] S.-H. Lin and J.-M. Ho. Discovering Informative Content Blocks from Web Documents. The 8th ACM SIGKDD, 2002.

[7] C. E. Shannon, A mathematical theory of communication. Bell System Technical Journal, 27:398-403, 1948.

[8] W3C DOM. Document Object Model (DOM). http://www.w3.org/DOM/.

Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03)
0-7695-1932-6/03 $17.00 © 2003 **IEEE**