Mining Web Informative Structures and Contents Based on Entropy Analysis

Hung-Yu Kao, Shian-Hua Lin*, Jan-Ming Ho+, Ming-Syan Chen

Electrical Engi. Dept.
National Taiwan University
Taipei, Taiwan, ROC
bobby@arbor.ee.ntu.edu.tw,
mschen@cc.ee.ntu.edu.tw

Dept. of Computer Science and Information Science*
National Chi Nan University
Nantou Hsien, Taiwan, ROC
shlin@csie.ncnu.edu.tw

Institute of Information Science+
Academia Sinica
Taipei, Taiwan, ROC
hoho@iis.sinica.edu.tw

Abstract

In this paper, we study the problem of mining the informative structure of a news Web site that consists of thousands of hyperlinked documents. We define the informative structure of a news Web site as a set of index pages (or referred to as TOC, i.e., table of contents, pages) and a set of article pages linked by these TOC pages. Based on the Hyperlink Induced Topics Search (HITS) algorithm, we propose an entropy-based analysis (LAMIS) mechanism for analyzing the entropy of anchor texts and links to eliminate the redundancy of the hyperlinked structure so that the complex structure of a Web site can be distilled. However, to increase the value and the accessibility of pages, most of the content sites tend to publish their pages with intra-site redundant information, such as navigation panels, advertisements, copy announcements, etc. To further eliminate such redundancy, we propose another mechanism, called InfoDiscoverer, which applies the distilled structure to identify sets of article pages. InfoDiscoverer also employs the entropy information to analyze the information measures of article sets and to extract informative content blocks from these sets. Our result is useful for search engines, information agents, and crawlers to index, extract and navigate significant information from a Web site. Experiments on several real news Web sites show that the precision and the recall of our approaches are much superior to those obtained by conventional methods in mining the informative structures of news Web sites. On the average, the augmented LAMIS leads to prominent performance improvement and increases the precision by a factor ranging from 122% to 257% when the desired recall falls between 0.5 and 1. In comparison with manual heuristics, the precision and the recall of InfoDiscoverer are greater than 0.956.

Keywords: Informative structure, link analysis, hubs and authorities, anchor text, entropy, information extraction.

1. Introduction

Recently, there has been explosive progress in the development of the World Wide Web. This progress creates numerous and various information contents published as HTML pages on the Internet. Furthermore, for the purpose of maintenance, flexibility and scalability of Web sites, Web publishing techniques are migrating from writing static pages to deploying dynamic application programs, which generate contents on requests based on predefined templates and contents stored in back-end databases. Most commercial Web sites, such as portal sites, search engines, e-commerce stores, news, etc., apply the dynamic technique to adapt diverse requests from numerous Web users. Such Web sites are called *systematic* Web sites in this paper. A news Web site that generates pages with daily hot news and archives historic news is a typical example of the systematic Web site.

Due to the evolution of automatic generation of Web pages, the number of Web pages grows explosively [9] [17]. However, there is a lot of redundant and irrelevant information on the Internet [8], such as contents of mirror sites or identical pages with different URLs [4][5]. We call this kind of redundancy *inter-site redundancy*. Also, a lot of redundant information exists within a Web site, especially in pages automatically generated by systematic Web sites. Such redundancy is referred to as *intra-site redundancy*. Examples of intra-site redundancy include company logos and texts, navigation panels, advertisements, catalogs of services, announcements of copyright and privacy policy. These contents are frequently texts or hyperlinks irrelevant to the meaning of the page, while said hyperlinks are used for easy access to other pages that are semantically irrelevant to the original page. In a systematic news Web site, adding such irrelevant links is convenient for users to browse other news articles with fewer clicks by following shortcuts in the page. However, these irrelevant links increase the difficulty for Web site analyzers, search engines and Web miners to perform their tasks. Those systems try to analyze, index and mine information from the whole site, including redundant and irrelevant information. However, the performance of these systems is unavoidably degraded by the redundant and irrelevant information.

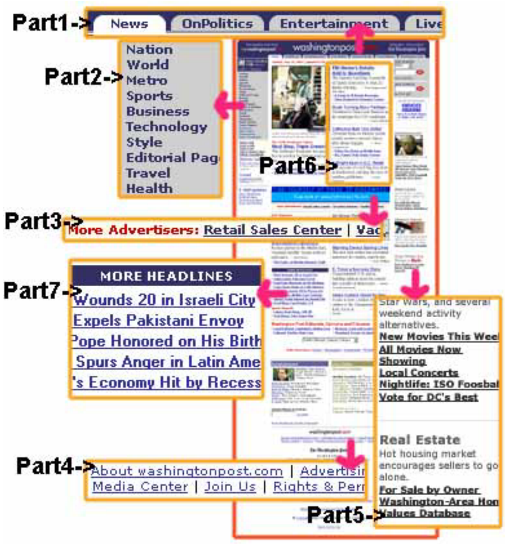


Figure 1: A sample page of news Web sites.

Consider the example in Figure 1. We divide the root page of WashingtonPost[1] into several parts with different styles and contents, i.e., (1) a banner with links "news", "OnPolitics", "Entertainment", "Live Online", etc. at the top, (2) a menu with 22 links of news categories on the left, (3) a banner with advertisement links, (4) general announcements about WashingtonPost, (5) a block with promoted hot news and advertisements, (6) a TOC block, and (7) a list with headline news. In this case, parts (1) and (2) are distributed among most pages in the site and are therefore redundant for users. We call these identical blocks *redundant content blocks*. However, they are still indexed by search engines. Such indexing induces an increase of the index size, and is useless for users and harmful for the quality of search results. Parts (3), (4) and (5) are irrelevant to the context of the page and are called *irrelevant content blocks*. These parts will make the topic of the page drift when terms in these parts are indexed. The last two parts, (6) and (7), draw more attention from users and are called *informative content blocks*,

[1] http://www.washingtonpost.com, a popular English news Web site.

in which users are able to read news articles via one click from anchors. For a user whose visiting is to read news, parts except for (6) and (7) are insignificant, since they are used for visual and traversal purposes. The following examples describe their impacts with more detail:

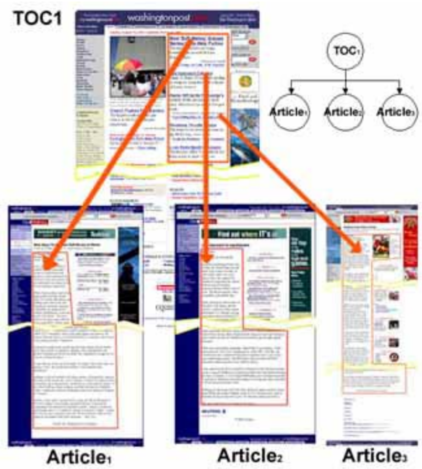


Figure 2: An example of the informative structure

***Example 1**:* After searching "game hardware tech jobs" in Google (http://www.google.com)[2], one of the most popular search engines, we found 20 pages of CNET (http://www.cnet.com, a news and product review Web site for computing and technology) in the top-100 results. However, none of these pages came from the Web pages categorized in CNET Job Seeker[3], which contains the desired information. There are matched query terms in redundant parts of pages among all these 20 pages, and however, three of these pages do not contain any matched terms in the informative parts of pages. The matched terms in redundant parts of a page will increase the rank of that page, even though they are usually ignored by users. Note that six of the 20 pages are ranked as top-16 and the page with the highest ranking does not contain any desirable information in its informative parts.

---

[2] The result is queried from www.google.com on February 5, 2002.
[3] CNET Job Seeker: http://dice.cnet.com/seeker.epl?rel_code=1&op=1.

Example 2*:* We estimate the redundant rate of news Web sites from our news search engine (NSE) that collects articles pages from several news Web sites in Taiwan[4]. According to the hand-coded rules, NSE merely crawls informative structures (TOC and article pages) and indexes the daily updated informative contents (article pages) from news Web sites. In a certain news Web site, one month of online news articles are reachable through their Web links. Based on these rules, it suffices to index only 5% of their pages and the index size of their informative parts is about 12.14% of the original page size. This implies that a typical search engine always indexes too much redundant information. Also, Web miners spend more effort in analyzing the whole site rather than focusing on the informative parts.

In a systematic Web site, we define a set of TOC (Table of Content) pages and a set of article pages linked by these TOC pages as the *informative structure* of a Web site. Furthermore, both kinds of pages are analyzed to extract informative content blocks. The informative structure of a Web site is therefore represented as a set of TOC blocks pointing to a set of article blocks. An example of the informative structure is illustrated in Figure 2. In the page TOC1, the content block enclosed by red line is the root of the structure and points to three article pages each containing a news article enclosed in a so called informative block. News agents or other applications may then access all the news article pages by traversing the informative structure. In this paper, we define the problem of mining informative structures from Web sites and present our solutions. We comment that the informative structure is useful in the following applications:

- Crawlers and Information Agents can focus on the informative structure to precisely and efficiently extract information for further analysis.

[4] http://nse.yam.com/. The news search engine collects news pages of fifteen news Web sites in Taiwan.

- Search Engines may index only the informative parts of an article page rather than indexing every page in the whole Web site. As a consequence, we are able to reduce the index size and increase the retrieval precision.
- Previous researches on Web miners (information extraction systems, e.g., WIEN [26], Stalker [30], IEPAD [16], and [35] expect the input Web pages to possess a high degree of regularity so that structured information, e.g., metadata, encoded in these pages can be retrieved. Informative structure is a key to automatically locate target informative content blocks containing useful repeated patterns in the whole site.

To illustrate the difficulty of the problem, we consider a real Web site in Figure 3, which shows a small graph of a Web site that contains only 270 pages and 1515 links. The graph looks like a messy "Web". The graph shown in Figure 4 is a subgraph (informative structure) extracted from the original graph in Figure 3 based on manually labeling TOC and article pages of the Web site. That is the answer set of the informative structure. This concise graph in Figure 4 only contains TOC pages and the corresponding article pages, which excludes 76.8% of the links and 6.7% of the pages from the original graph. Figure 4 consists of several trees whose root nodes are TOC pages. In this paper, we propose methods to reduce complicated Web site structures such as the one in Figure 3 into concise informative structures as in Figure 4.

Explicitly, we propose in the paper mechanisms to automatically discover informative structure of a Web site and contents of pages to reduce intra-site redundancy. We also present an entropy-based analysis to estimate the information quality of links and content blocks. This new entropy measure is used in the approaches for discovering informative structure. In the rest of the paper, we first present a literature survey on related studies of the paper in Section 2. In Section 3, we present a formal model on the problem and develop our entropy-based analysis. Following the system design and implementation in Section 4, we perform several experiments on real datasets to evaluate our methods in Section 5. Finally, we conclude the paper and describe the direction of future research in Section 6.

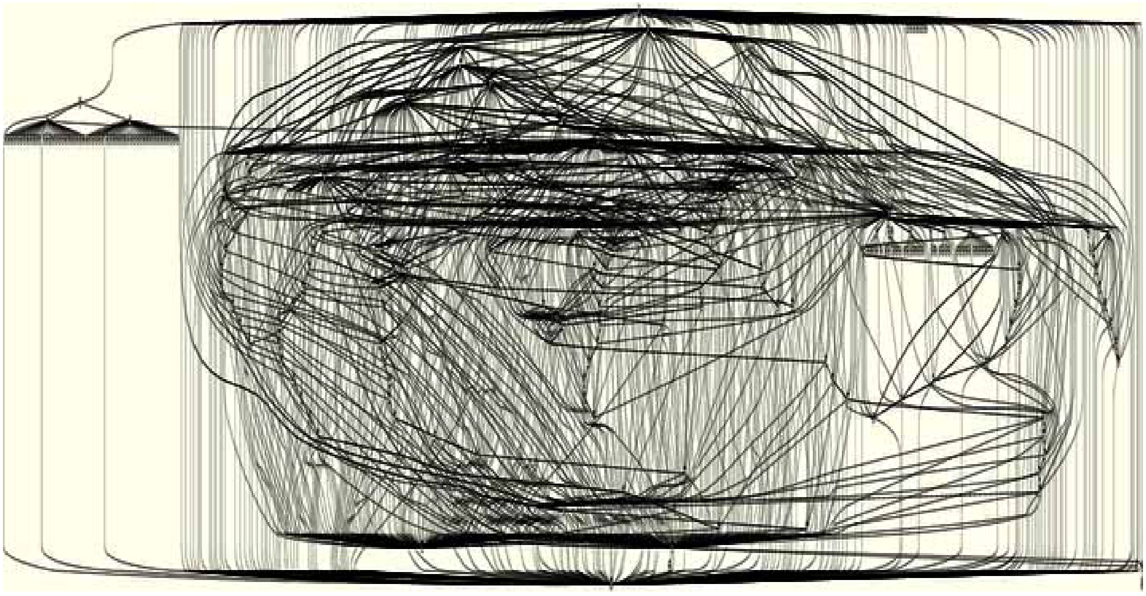


Figure 3: The linked graph of CDN (www.cdn.com.tw). Page numbers N = 270 and total links L = 1515.

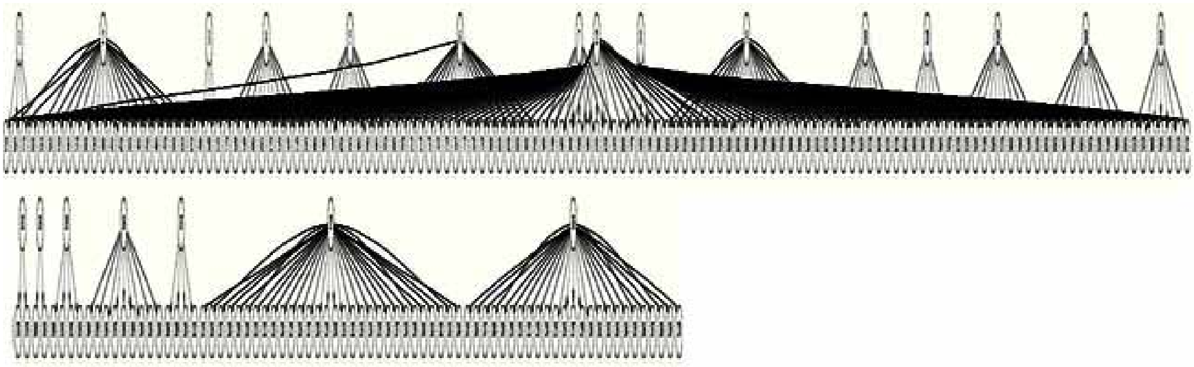


Figure 4: The reduced subgraph contains TOC and article pages. N = 252 and L = 350.

2. Related Work

The Hyperlink Induced Topics Search (HITS) algorithm [25] and Google's PageRank [7] are widely applied to analyze the structure of the Web. HITS estimates the authority and hub values of hyperlinked pages in the Web, and Google merely ranks pages. Both methods are applied to ranking the search result. Based on mutual reinforcement relationship, HITS provides an innovative methodology for Web searching and topics distillation. According to the definition in [25], a Web page is an authority on a topic if it provides good information, and is a hub if it provides links to good authorities. In recent research work on link analysis of hyperlinked documents, HITS is applied to the research area of topic distillation and several kinds of link weights are proposed to enhance the significance of links in hyperlinked documents. In the Clever system [14], weights tuned empirically are added to distinguish same-site links and others. In [3], the similarity between the document and the linked document is taken as the link weight for analysis. Another study that uses the similarity between the surrounding text of a link and the linked document to determine the link weight is conducted in [13]. Considering the distribution of terms in documents, Chakrabarti [11] combines the TFIDF-weighted model and micro-hub to represent the significance of links in regions with information needed.

Intuitively, HITS and its related methods applied to the topic distillation are useful in the analysis of Web informative structures. However, the topic distillation is different from the informative structure mining in several aspects:

- The former distills hubs and authorities from a set of pages retrieved from search engines with a given query. These pages are not restricted to be published from the same Web site. However, the latter mines the informative structure from all pages of a given Web site.
- With different targets of data sets, studies of topic distillation usually omit intra-links and nepotistic links to perform the mutual reinforcement between sites. However, these links are important while the link analysis is focused on a Web site.
- Most adaptive topic distillation algorithms based on HITS take the relationship between queries and documents into consideration. However, these algorithms do not work well on mining informative structures because of the absent of a target query. Furthermore, as described in [12][27], the link analysis algorithms, e.g. HITS, are vulnerable to the effect of nepotistic clique attack and Tightly-Knit Community (TKC). The effects will be more significant for mining informative

structures of Web sites since we observed that nepotistic links and cliques appear more frequently in a Web site [24].

Based on mining the informative structure of a Web site, the complex structure is reduced to a concise one. However, if we look into pages of the structure, many redundant content blocks are not meaningful for the pages content. In [20] and [21], studies provide learning mechanisms to recognize advertisements and redundant/irrelevant links of Web pages. However, these methods need to build the training data first and related domain knowledge must be included to extract features for generation of classification rules. Therefore, both methods are difficult to be applied to automatically extract the informative structures of systematic Web sites.

Studies of Information Extraction (IE) [21][22][26][35] aim to mine structure values (metadata) of pages from Web sites. Although being able to extract valuable metadata from pages, most of these IE systems need labor-intensive work. Cardie [10] defines five pipelined processes for an IE system: tokenization and tagging, sentence analysis, extraction, merging, and template generation. Machine learning is usually applied to learn, generalize, and generate rules in the last three processes based on manually generated domain-specific knowledge such as concept dictionaries and templates. Training instances applied to learning processes are also artificially selected and labeled. For example, in Wrapper induction [26], the author manually defines six wrapper classes, which consist of knowledge to extract data by recognizing delimiters to match one or more of the classes. The richer a wrapper class, the more likely it will work with any new site [15]. SoftMealy [22] provides a GUI that allows a user to open a Web site, define the attributes and label the tuples in the Web page. The common disadvantages of IE systems are the cost of templates, domain-dependent knowledge, or annotations of corpora generated by hand. This is the very reason that these systems are merely applied to specific Web applications, which extract the structural information from pages of specific Web sites or pages generated by CGI.

Consequently, IE systems are not scalable and therefore cannot be applied to resolve the problem of redundant content blocks in pages.

3. The Entropy-Based Analysis

In this section, we propose an entropy-based analysis to remedy the deficit of HITS-related algorithms. In light of this analysis, we devise two mechanisms to mine the informative structure of a Web site. We first develop a mechanism of analyzing the entropy of anchor texts and links, namely Link Analysis of Mining Informative Structure (LAMIS), to reduce a complex Web site to a distilled concise Web structure. Then, we devise another mechanism on analyzing the entropy of content blocks, namely InfoDiscoverer, to identify informative (significant) content blocks from pages in the concise structure.

3.1 LAMIS – Analyzing Entropy of Anchor Texts and Links

Given an entrance of a Web site, HITS is applied to measure hub and authority values of all pages. Intuitively, a TOC page is expected to have a high hub value, and an article page is to have a high authority value. That is, HITS is designed to provide a reasonable solution to mining the informative structures of news Web sites. However, the existence of redundant and irrelevant links usually causes the phenomenon of topic drift in pages published in systematic Web sites. Therefore, the analysis of authority and hub cannot solely depend on the linked structures with the page granularity. We apply the entropy to measure the significance of anchor texts and page content while using HITS on the link analysis. For example, when the link entropy is applied to shrink the CDN graph shown in Figure 3, the thresholds of entropy values, 0.8 and 0.4, reduce that graph to those shown in Figure 5 and Figure 6, respectively. In comparison with the answer set shown in Figure 4, Figure 6 reduced by the threshold 0.4 indeed approaches to the optimal solution. Explicitly, the objective of our first mechanism, LAMIS, is to reduce a complex graph to an optimal subgraph that represents the informative TOC and article structure. The detail is described in the following sections.

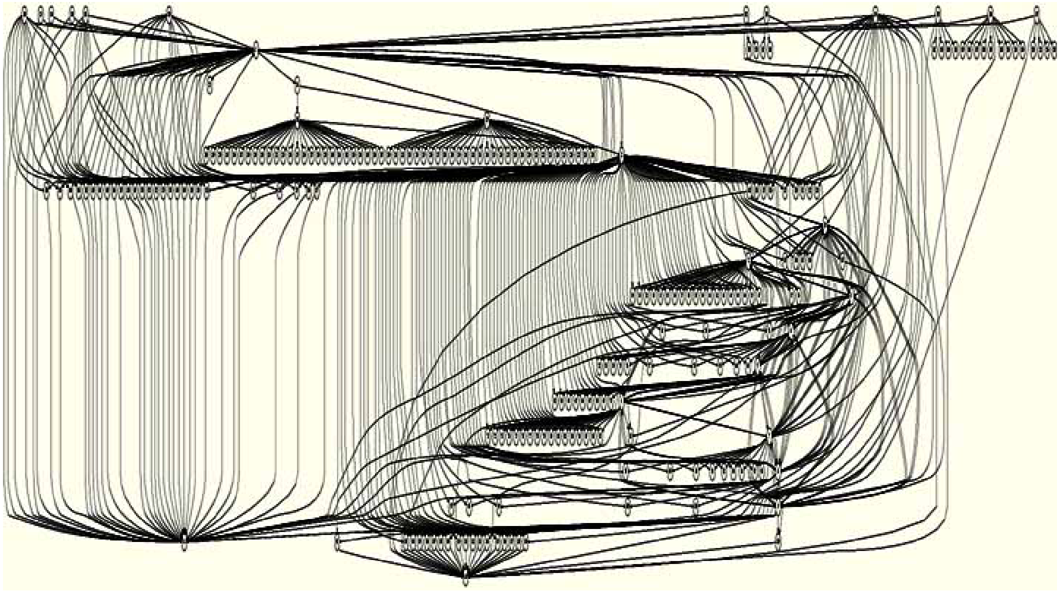


Figure 5: The graph of CDN: links with entropy values smaller than 0.8 [N=241, L=569].

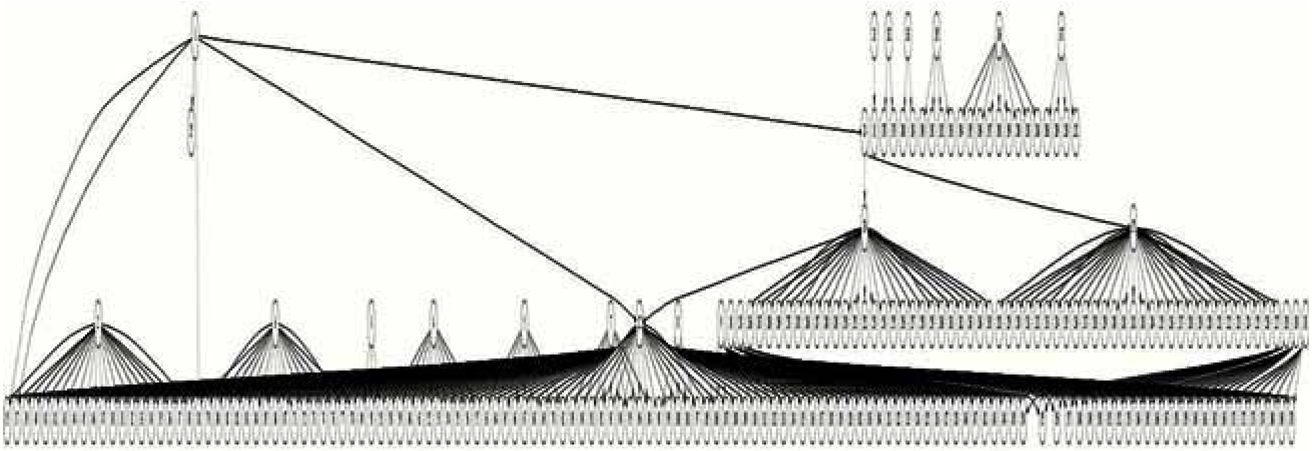


Figure 6: The graph of CDN: links with entropy values smaller than 0.4 [N=236, L=353].

3.2 InfoDiscoverer – Analyzing Entropy of Content Blocks

The entropy-based analysis can further be applied to analyze page content blocks and discover informative contents from article pages clustered by LAMIS. Our second mechanism, *InfoDiscoverer*, standing for "discovering informative content blocks," is designed to analyze the information measure of content blocks (in a page set) and dynamically select the entropy-threshold to classify a page's blocks into either informative or redundant. By partitioning a page into blocks based on HTML tags, InfoDiscoverer calculates entropy values of features (i.e., terms, keywords, or phrases) according to the probability distribution of features in the page set. Entropy values of content blocks are derived from their features. Based on the answer set generated from 13 manually tagged news Web sites with a total of 26,518 Web pages, experiments show that both recall and precision rates are greater than 0.956.

Consequently, LAMIS applies the link entropy to discover the informative structure and InfoDiscoverer employs the content entropy to determine the content property, informative or redundant. In the following, we describe the derivation of content and link entropy values. Then, we develop mechanisms to apply link entropy values to enhance the link analysis and mine the informative structure.

3.3 The Entropy of Content and Link and Enhanced Link Analysis

The text is an important clue for users to determine the meaning of a page. Therefore, we extract features (terms) from the page text to represent its corresponding significance. For the same reason, the anchor text is an important clue for users to click the link. Therefore, we also extract features (terms) from an anchor text to represent the significance of the link. In this paper, a term corresponds to a meaningful keyword. The idea is that features frequently appearing in most pages are redundant and carry less information to users. In contrast, features appearing in fewer pages are more informative. That is, we can apply the probability distribution of terms in pages to estimate the information measures of features. First, we calculate the feature entropy and deduce the entropy values of content blocks and links.

Given a set of HTML pages or a Web site, we can parse all pages into content blocks and links based on HTML tags. Also, texts appearing in contents and links can be parsed into a set of terms with corresponding blocks and links. In the page set, those features form a feature-document matrix with weights in matrix entries. The feature weight is calculated based on the calculation of feature frequency in the page set [29]. Then, Shannon's information entropy [34] is applied to calculate the feature entropy based on the feature-document matrix. By definition, the entropy E can be expressed as $-\sum_{i=1}^{n} p_i \log p_i$, where p_i is the probability of $event_i$ and n is number of events. By normalizing the weight of a feature to be [0, 1], the entropy of feature (term) T_i is:

$$E(T_i) = -\sum_{j=1}^{n} w_{ij} \log_2 w_{ij} ,$$

in which w_{ij} is the value of normalized feature frequency in the page set. To normalize the entropy to the range [0, 1], the base of the logarithm is chosen to be the number of pages, and the previous equation becomes:

$$E(T_i) = -\sum_{j=1}^{n} w_{ij} \log_n w_{ij} , \textit{ where } n = |D|, D \textit{ is the set of pages}.$$

Entropy values of content blocks and links are derived from the average of their features entropies. For the example of calculating the entropy of a content block, intuitively, feature entropies contribute to the semantic measure of a content block that owns these features. I.e. the entropy value of a content block is the summation of its features entropies:

$$H(CB_i) = \sum_{j=1}^{k} H(F_j), \text{ where } F_j \text{ is a feature of } CB_i \text{ with } k \text{ features}.$$

Since content blocks contain different numbers of features, the equation is normalized as:

$$H(CB_i) = \frac{\sum_{j=1}^{k} H(F_j)}{k} .$$

That is, the entropy of a content block, H(CB), is the average of all feature entropies in the block. The link entropy is measured analogously.

In the link graph of a Web site *G=(V, E)*, HITS algorithm computes two scores for each node *v* in *V*, i.e., the hub score *H(v)* and the authority score *A(v)*. In our approach, we incorporate entropy values of links as link weights to present the significance of links in a Web site structure. Therefore, the original HITS algorithm is modified as follows:

$$A(v) = \sum_{(u,v)\in E} H(u) * \alpha_{uv} \text{ and } H(v) = \sum_{(v,u)\in E} A(u) * \alpha_{uv},$$

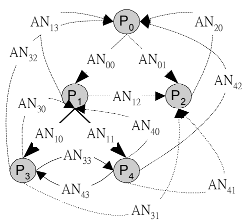
where α_{uv} is the weight of the link from u to v, denoted AN_{uv}. According to the definition of entropy of a link, α_{uv} is defined as follows:

$$\alpha_{uv} = 1 - E(AN_{uv}).$$

It can be seen that the more information a link carries, the larger the link weight is, i.e., the lower the link entropy is.

Moreover, the SALSA proposed in [27] is designed to resist effects of TKC and cliques and we also apply entropy weights on SALSA to remedy the effects similarly. The improvement will be empirically evaluated by our experimental studies later.

3.4 An Illustrated Example



	P_0	P_1	P_2	P_3	P_4
P_0	\	1	1	0	0
P_1	1	\	0	1	1
P_2	1	0	\	0	0
P_3	1	1	1	\	0
P_4	1	1	1	0	\

Adjacent Matrix

P_0

AN_{00}: "hot news"
AN_{01}: "sales"

P_1

AN_{10}: "sports"
AN_{11}: "business"
AN_{12}: "sales"
AN_{13}: "home"

P_2

AN_{20}: "home"

P_3

AN_{30}: "hot news"
AN_{31}: "sales"
AN_{32}: "home"
AN_{33}: "business news"

P_4

AN_{40}: "hot news"
AN_{41}: "sales"
AN_{42}: "home"
AN_{43}: "sports news"

	P_0	P_1	P_2	P_3	P_4
T_0 (hot)	1	0	0	1	1
T_1 (news)	1	0	0	2	2
T_2 (sales)	1	1	0	1	1
T_3 (sports)	0	0	1	0	1
T_4 (business)	0	0	1	1	0
T_5 (home)	0	1	1	1	1

TD Matrix

Figure 7 A simple Web site, |D|=5.

Considering the example of a simple news Web site shown in Figure 7, page P_0 is the homepage of the Web site. Page P_1 is the TOC page with two anchors linking to news article pages, P_3 and P_4. Page P_2 is an advertisement page linked by the other four pages. Most pages contain anchor texts, i.e., "home", "hot news", and "sales", linking to P_0, P_1, and P_2 respectively. P_3 and P_4 have anchors linking to each other. It is regarded as a cross-reference to present the related news. The Web site structure is also widely used in commercial Web sites. Based on the terms in each page, the feature entropy of the Web site is calculated as below:

$$E(T_0) = -\sum_{j=1}^{3} \frac{1}{3} \log_5 \frac{1}{3} = 0.682,$$

$$E(T_1) = -\sum_{j=1}^{2} \frac{2}{5} \log_5 \frac{2}{5} - \frac{1}{5} \log_5 \frac{1}{5} = 0.655,$$

$$E(T_2) = E(T_5) = -\sum_{j=1}^{4} \frac{1}{4} \log_5 \frac{1}{4} = 0.861, \textit{ and}$$

$$E(T_3) = E(T_4) = -\sum_{j=1}^{2} \frac{1}{2} \log_5 \frac{1}{2} = 0.430.$$

The entropy values of links derived from feature entropies are listed in Table 1. According to the definition of entropy, the most informative links are AN_{10} and AN_{11}, which link to article pages P_3 and P_4 from TOC page P_1.

Table 1: Entropy values of links shown in Figure 7.

P_0		P_1				P_2	P_3				P_4			
AN_{00}	AN_{01}	AN_{10}	AN_{11}	AN_{12}	AN_{13}	AN_{20}	AN_{30}	AN_{31}	AN_{32}	AN_{33}	AN_{40}	AN_{41}	AN_{42}	AN_{43}
0.669	0.861	**0.430**	**0.430**	0.861	0.861	0.861	0.669	0.861	0.861	0.543	0.669	0.861	0.861	0.543

Table 2: Results of link analysis of HITS and Entropy-based HITS.

Method	HITS		Entropy-based HITS	
	Authority	Hub	Authority	Hub
P_0	0.535	0.297	0.229	0.142
P_1	0.419	0.524	0.338	**0.756**
P_2	**0.576**	0.160	0.244	0.031
P_3	0.321	**0.553**	**0.622**	0.451
P_4	0.321	**0.553**	**0.622**	0.451

Based on the link entropy, we use our entropy-based HITS algorithm to calculate values of hub and authority. In comparison with HITS shown in Table 2, hub and authority values are obtained after 10 iterations. P_1 is ranked as the top-1 hub page by the entropy-based HITS, and P_3 and P_4 are ranked with the highest authority. However, HITS ranks the advertisement page (P_2) as the best authoritative page, and news article pages (P_3 and P_4) as good hub ones. It can be seen that the link entropy is effective to enhance the link significance in link analysis algorithms.

4. The System Design

In this section, we will describe the design of LAMIS and InfoDiscoverer. The LAMIS system is designed to explore hyperlink information to extract and identify the hub (TOC) and authority (article) pages. InfoDiscoverer then measures the entropy values of content blocks among clustered pages (a set of article or TOC pages) and dynamically select the entropy-threshold to extract informative content blocks. Given an entrance URL of a Web site, without manual inventions and prior knowledge about the Web site, LAMIS and InfoDiscoverer are capable of crawling all pages, analyzing the structure, and extracting the informative structures and content blocks of the site. In this section, we describe the system architecture and processes of LAMIS and InfoDiscoverer.

4.1 The System Architecture

Our Web mining system, shown in Figure 8, consists of three parts:

(1) Web Crawler which crawls pages, parses them into blocks, and builds the link graph of the Web site.

(2) Feature Extractor which extracts features (terms), in-degrees, out-degrees, text lengths, and links as metadata of pages. Feature Extractor also calculates entropy values of features to derive entropies of links and content blocks.

(3) Informative structure mining module which is composed of LAMIS and InfoDiscoverer.

First, a starting URL of a site is given to Crawler. In our system, we can assign crawl depths to different paths (sub-trees) of the site. Once a site has been crawled, a page is represented by some content blocks, and a site structure is built, and terms appearing in all pages are extracted. Terms appearing in links or content blocks are recorded for the subsequent estimation of feature entropy. As we know, extracting English terms is relatively simple. Applying stemming algorithms and removing stop words based on a stop-list, English keywords (terms) can be extracted [33]. Extracting terms used in oriental languages is in general more difficult because of the lack of separators in these languages. However, many studies have applied statistical approaches to extracting keywords of oriental languages [18]. In our system, we use an algorithm to extract keywords from Chinese sentences based on a Chinese term base generated via collecting hot queries, excluding stop words, from our search engine[5]. After extracting features, the system maintains a feature-document matrix to represent the feature frequency corresponding to each document. According to the matrix, we can calculate the entropy values of features and derive links and content blocks entropies to be used as inputs to LAMIS and InfoDiscoverer.

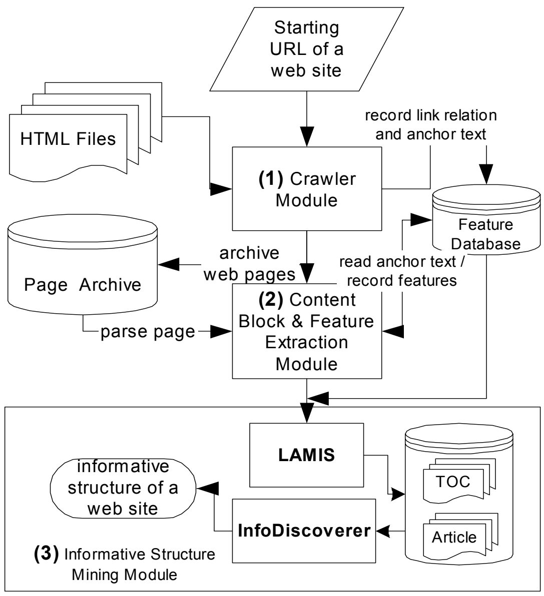


Figure 8: The system architecture.

[5] The searching service is a project sponsored by Yam, (http://www.yam.com/).

4.2 LAMIS: Mining Informative Structures

To employ the link entropy in enhancing the link analysis algorithms, the most important issue is to prove the revised algorithm is convergent. In HITS, hub and authority will converge to the principal eigenvector of the link matrix [25]. Also, the weighted HITS algorithm is converged, if the multiplication of both weight matrices has no negative entry values [3]. In LAMIS, the weighting factor α_{uv}, is bounded in [0,1], and we use the same weight matrix in hub and authority. Therefore, LAMIS will be converged after constant iterations. In our experiments on various sites, 91.7% of the hub values will converge to zero after 10 iterations. Figure 9 shows one of the experimental results in which hub values of all pages are sorted in descending order. It can be seen that they decrease sharply and most of them become zero. From our observation, TOC pages tend to hold high hub values, and we use the top-N threshold to extract TOC pages from the ranked list of the whole page sets.

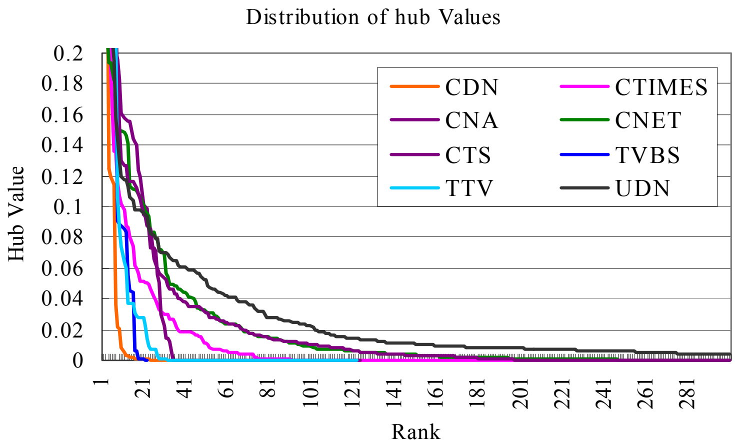


Figure 9: Distribution of hub values.

During the iteration of HITS, the converged curves are slightly undulated as shown in Figure 10. This phenomenon is due to the effect of multiple propagation paths of mutual reinforcement relationships in HITS algorithm and is called *non-uniqueness* in [30]. Consider Figure 11 for example. It is seen that two propagation paths are independent, and the two converged authority and hub sets of one page, i.e.,

(A_{2k}, H_{2k}) and (A_{2k+1}, H_{2k+1}), will hence be generated. While the authority and hub sets of all pages are considered, the two independent sets must be combined alternately. In Figure 11, if these values converge at iteration 3, two authority and hub sets are $\{(A_{a3}, H_{a3}), (A_{b2}, H_{b2}), (A_{c3}, H_{c3})\}$ and $\{(A_{a2}, H_{a2}), (A_{b3}, H_{b3}), (A_{c2}, H_{c2})\}$. In general cases, these two independent sets will converge to the same one and we may select one of them to be used subsequently.

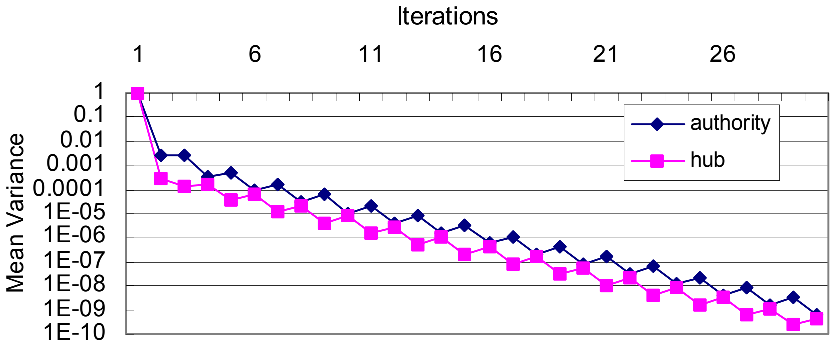


Figure 10: The convergence of authority and hub values.

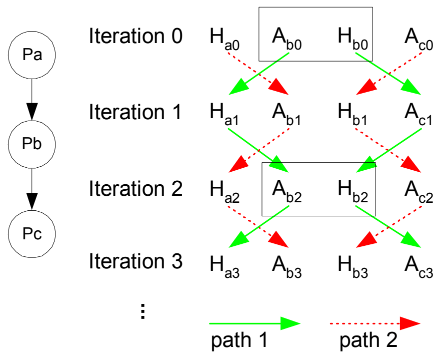


Figure 11: Two independent propagation paths of mutual reinforcement relationships in HITS.

4.3 InfoDiscoverer: Mining Informative Contents

After identifying TOC pages from a Web site, we can find article pages by following links in TOC pages. However, redundant and irrelevant links in article pages are not easy to be discovered. Therefore, we apply InfoDiscoverer to extract informative content blocks of the set of TOC pages. Article pages are defined as pages linked by anchors appearing in informative blocks of a TOC page. Also, these article pages form a new data set from which InfoDiscoverer extracts informative blocks as the meaningful content article pages.

Given a set of TOC or article pages, InfoDiscoverer classifies content blocks into two categories, redundant and informative, based on the entropy values of content blocks as follows:

- If the entropy of a content block is higher than a defined threshold or close to 1, the block is absolutely redundant since most of the block's features appear in every page.
- If the entropy of a content block is less than a defined threshold, the block is informative because features of the page are distinguishable from others, i.e., these features of the page seldom appear in other pages.

The threshold is not easy to determine since it would vary for different page sets. If the higher threshold is chosen, the higher recall rate is expected. However, the precision rate may become lower. To get a balanced recall-precision rate, we apply the greedy approach to dynamically determine the threshold for different page sets. If the threshold is increased, more informative features (in informative content blocks) will also be included. The basic idea of the greedy approach is described as the following heuristic.

- *Starting the entropy-threshold from 0 to 1.0 with an interval such as 0.1, increasing threshold value will include more features since more content blocks are probably included. If the increase of the threshold never includes more features, the boundary between informative and redundant blocks is reached.*

5. Experiments and Evaluation

In this section, we describe experiments on several news Web sites to evaluate the performance and improvement of our approaches. We first describe the datasets used in experiments. Then, we describe the evaluation of LAMIS and InfoDiscoverer and assess the performance of both methods.

5.1 The Datasets

In our experiments, the datasets[6] contain fourteen Chinese and five English news Web sites as described in Table 3. All of these news sites provide real-time news and historical news browsing services including several domains: politics, finance, sports, life, international issues, entertainment, health, cultures, etc. In our experiments, the crawl depth is set to 3, and after pages have been crawled, the domain experts[7] inspect the content of each page in Chinese news Web sites. They labeled these pages as TOC or article to build the answer set of datasets used in the following experiments. We found that the percentages of TOC pages vary among sites, i.e., different sites have different policies to organize their site structure. In fact, several sites have particular information structures. The diversity of information structures in datasets demonstrates the general applicability of LAMIS.

5.2 Evaluation on LAMIS

After extracting features from crawled pages, we compute the entropy values of content blocks and anchors. We found that entropy values of 71.6% of links are larger than 0.8, and they are probably redundant. As we expect, they appear in links or contents of navigation panels, advertisements, and copyright announcements, etc. We first compare the performances of three link analysis algorithms: HITS, SALSA, and LAMIS. Basically, SALSA is the HITS algorithm with link normalization (LN) [6] and is therefore symbolized by HITS-LN. Similarly, the link normalization can be employed in LAMIS

[6] Pages of Web sites in datasets are crawled at 2001/12/27 and 2002/4/11. The datasets can be retrieved in our research site http://kp06.iis.sinica.edu.tw/isd/index.html.

[7] The domain experts are the site managers of Yam News Search Engine (NSE, http://news.yam.com).

Table 3: Datasets and related information

Site Abbr.	URL of Root	Total pages	TOC pages	Links	Content Blocks
CDN	www.cdn.com.tw/welcome.htm	261	25	1,339	892
CTIMES	news.chinatimes.com/	3,747	79	26,848	79,077
CNA	www.cna.com.tw/	1,400	33	5,849	14,544
CNET	taiwan.cnet.com/news/	4,331	78	25,844	15,912
CTS	www.cts.com.tw/	1,316	31	8,915	16,149
TVBS	www.tvbs.com.tw/code/tvbsnews/index.asp	740	13	3,530	5,937
TTV	www.ttv.com.tw/HomeV2/default.htm	861	22	3,301	4,990
UDN	udnnews.com/NEWS/	4,676	252	34,882	84,411
CNN	www.cnn.com	626	N/A*	21,276	11,643
WP	www.washingtonpost.com	1,301	N/A	10,367	8,203
LATIMES	www.latimes.com	1,119	N/A	25,069	8,720
CSMONITOR	www.csmonitor.com	3,618	N/A	31,972	14,260
DISPATCH	www.dispatch.com	603	N/A	711	5,862
ITHOME	www.ithome.com.tw/News/Investment/	202	N/A#	N/A	N/A
ET	www.ettoday.com.tw/life/	159	N/A	N/A	N/A
FTV	www.ftv.com.tw/	794	N/A	N/A	N/A
TSS	www.tssdnews.com.tw/cgi-bin/news_sub/	123	N/A	N/A	N/A
CTV	www.chinatv.com.tw	3,597	N/A	N/A	N/A
TTIMES	www.ttimes.com.tw	1,966	N/A	N/A	N/A

*: We only consider top-20 precision in experiments of English Web site. Hence, we do not find all TOC pages in English Web sites.

#: Datasets from ITHOME to TTIMES are only appended for the experiment of informative content block discovering. The last three columns are omitted.

to become LAMIS-LN. We also compare these algorithms with a simple heuristic algorithm Out-link (***OL***), which ranks pages according to their counts of out-link rather than hub values. The idea of algorithm OL comes from that TOC pages have more links than article pages in general. Therefore, these methods are denoted by: OL, HITS, HITS-LN (SALSA), LAMIS, and LAMIS-LN. Based on the datasets (with answer sets) of Table 3, we conduct experiments to evaluate the performance of the previous five methods. By ranking pages according to their hub values or out-link counts, we examine the precision at 11 standard recall levels [2] for each site. Figure 12 shows the precision histograms of these methods based on the average precision rates of Web sites in datasets. We observe that LAMIS-LN emerges as the winner, followed by heuristic OL. HITS does not work well in these datasets.

Note, however, that these methods do not suffice to simultaneously render high recall and precision rates. In view of this, we shall devise some techniques to enhance LAMIS in the following.

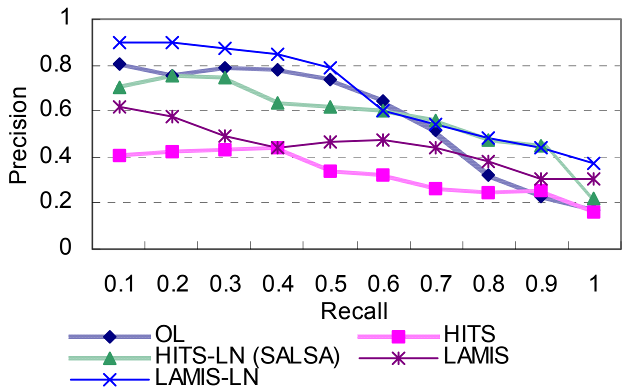


Figure 12: The effect of weights on link analysis.

5.2.1 Augmented Features of LAMIS

After investigating the raw pages in these datasets, we found that informative links are merely located in one or several informative content blocks. In previous experiments, we only consider hub values. The authority value is probably a compensation of the hub. Considering the anchor texts of links, redundant links are usually used for presentation purpose and their text lengths are thus usually very short. However, informative links tend to describe the title of the linked pages for the readable purpose. Therefore, we also consider the length of anchor text and consequently propose the following techniques to enhance our methods:

- Page mode (PA) vs. Content block mode (CB),
- Hybrid ranking (HR) of authority and hub, and
- Anchor text length of links, which is linear to the number (weight) of term counts (TW).

Page Mode vs. Content Block Mode

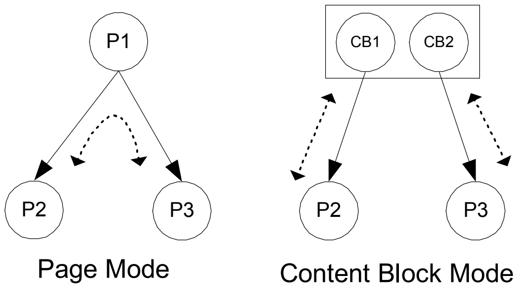


Figure 13: Propagations of mutual reinforcement on different modes

In Figure 13, we can see the difference of mutual reinforcement propagation between the page mode and the content block mode. In the page mode, authority of P2, i.e., A_{p2}, can affect the value A_{p3} through hub of P1, H_{p1}. If P2 is authoritative, A_{p3} will also be promoted, even though it is not an authoritative page. In the content block mode, we divide P1 into two blocks, one contains a link to P2, and the other contains a link to P3. They are treated as separate nodes. Hence, the propagation of high authority of P2 will now be terminated at CB1 and P2 will not be incorrectly promoted. In our approach, blocks in the content block mode are extracted based on the <TABLE> HTML tag.

Hybrid Ranking (HR) of Authority and Hub

When the contexts of pages are complex, pages may contain more hybrid characteristics of hub and authority. To reduce the effect of hybridization of hubs and authorities in a page, we take into consideration the authority and use the hybrid ranking of hubs and authority. The idea is motivated by the observation that TOC pages hold not only the higher hub values, but also the lower authority values. To capture this notion, by assuming that the hub is inversely proportional to the authority, we employ the following equation in our experiments,

$$Rank = hub - n * authority,$$

where n is a Web site dependent coefficient with its value determined by $\log_2(\frac{\textit{the number of links in the Web site}}{1000})$.

Applying Anchor Text Length to Reduce the Effect of Redundant Links

In our definition, redundant links are those links associated with navigation panels, advertisements, and others not relevant to the topic of the page. According to the information theory, terms in these links usually hold high entropy values, so that these links are less weighted in link analyses. In TOC pages, informative links are usually described by a sentence to summarize the content of the linked page. For example, the anchor text of a TOC page's link is usually the title of the linked article page. However, a readable sentence frequently consists of keywords and stop words, and the link entropy is therefore diluted. Therefore, the length of anchor text is taken into the consideration to enhance the performance of LAMIS. The anchor length is linear to the number of terms extracted from the anchor text; we define the term weight (TW) to evaluate the effect of the anchor text length:

$$\alpha_{uv} = \alpha_{uv} * (1 + \log_{10}(term\ count)) .$$

This equation means that if the number of terms in an anchor is 10, the weight of the link is doubled.

5.2.2 Performance Evaluation on LAMIS with Augmented Features

Considering augmented features mentioned previously, we have many optional combinations of methods and experiments. For a typical example, LAMIS-LN-CB-HR-TW means we integrate the "link normalization", "hybrid ranking", and "term weight of anchor text length" to LAMIS by analyzing authority and hub in "content block mode". In Table 4, we show some more experiments for combinations of these methods based on the R-Precision, i.e., the equal point of recall and precision. The average R-Precision, 0.82, of LAMIS-LN-CB-HR-TW (the optimal LAMIS) is the best result. The optimal LAMIS is ranked first for R-Precision in six of eight Chinese datasets and ranked second in the other two. As compared to HITS, LAMIS-LN-CB-HR-TW improves the R-Precision by a factor of 2.03.

Table 4: R-Precision of all experiments

R-Precision	CDN	CTIMES	CNA	CNET	CTS	TVBS	TTV	UDN	AVG.
outlinks	0.84	0.67	0.36	0.44	0.42	0.77	0.64	0.43	0.57
HITS	0.48	0.63	0.94	0.03	0.03	0.01	0.05	0.02	0.27
HITS-LN (SALSA)	0.92	0.77	0.97	0.42	0.29	0.77	0.36	0.25	0.59
LAMIS	0.88	0.63	0.30	0.33	0.16	0.92	0.05	0.06	0.42
LAMIS-LN	**0.96**	0.77	0.52	0.55	0.32	**1.00**	0.50	0.53	0.64
CB-HITS	0.48	0.22	0.24	0.46	0.52	0.01	0.36	0.28	0.32
CB-HITS-LN	**0.96**	0.33	0.30	0.45	0.39	0.69	0.23	0.56	0.49
LAMIS-CB	0.88	0.49	0.12	0.31	0.13	0.92	0.18	0.09	0.39
LAMIS-LN-CB	0.92	0.95	**0.97**	0.51	0.26	**1.00**	**0.86**	0.53	0.75
LAMIS-LN-CB-HR	**0.96**	0.95	0.94	**0.62**	0.29	0.85	**0.86**	0.68	0.77
LAMIS-LN-CB-HR-TW	**0.96**	**0.98**	**0.97**	0.58	**0.58**	0.85	**0.86**	**0.77**	**0.82**

OL: rank by number of outlinks in a page, **PA**: Page mode, **CB**: Content block mode, **HITS**: Kleinberg's HITS, LN: Link normalization, HITS-LN=**SALSA**: the stochastic approach for link-structure analysis, , **AEN**: weighted by anchor text entropy, **HR**: hybrid ranking, **TW**: term count weight

We also select the major methods and draw the graph from the experimental result. It can be seen that LAMIS-LN-CB-HR-TW outperforms others in all nine Chinese news Web sites in Figure 14.

To evaluate our methods in English Web sites, we conduct several experiments on English news Web sites and compare the top-20 precision rates as shown in Figure 15. We note that same as in Chinese news Web sites, LAMIS-LN-CB-HR-TW still performs very well, indicating the robustness of this method.

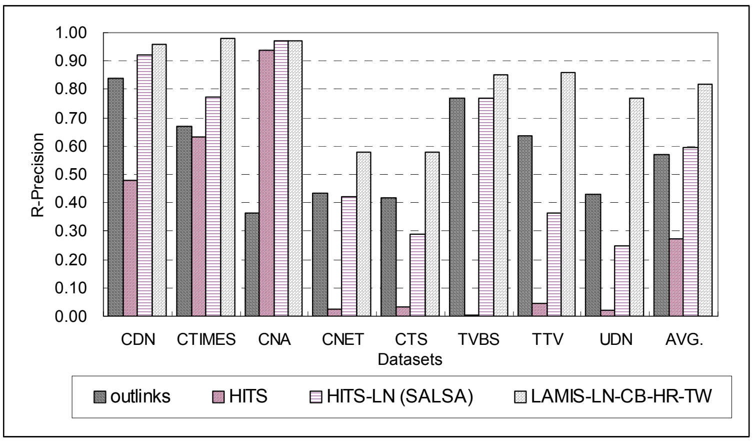


Figure 14: R-Precision improvement of augmented LAMIS.

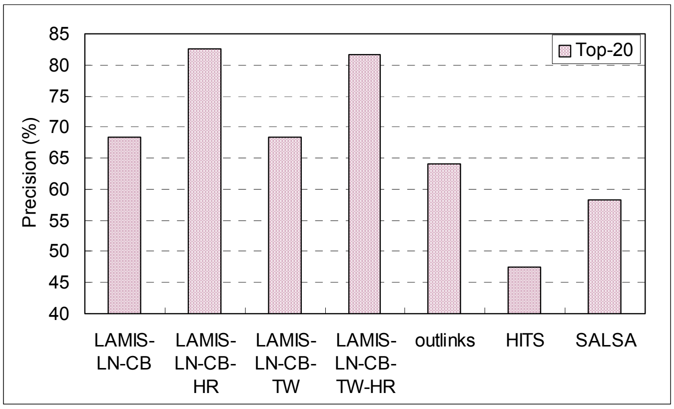


Figure 15: Top-20[8] precision diagrams of English news Web sites.

5.3 Evaluation on InfoDiscoverer

Table 5: News sites with tabular pages.

Site	Site+Path	Page set	Pages	Opt. Entropy	Recall	Precision
IThome	http://www.ithome.com.tw/News/Investment/	Net. Investment	202	0.7	0.957	1.000
ET	http://www.ettoday.com.tw/life/	Life	159	0.2	1.000	0.979
FTV	http://www.ftv.com.tw/	Taiwan News	794	0.4	1.000	1.000
CNet	http://taiwan.cnet.com.tw/investor/news/	Investment	499	0.5	0.956	1.000
TSS	http://www.tssdnews.com.tw/cgi-bin/news_sub/	Supplement	123	0.3	0.989	1.000
CDN	http://www.cdn.com.tw/daily/	Misc. News	1,305	0.5	1.000	1.000
TVBS	http://www.tvbs.com.tw/code/tvbsnews/daily/	Daily News	9,943	0.1	1.000	1.000
CTV	http://www.chinatv.com.tw/	Taiwan News	3,597	0.2	1.000	1.000
CAN	http://www.cna.com.tw/cgi-bin/readcipt77.cgi?a1&0	Headlines	5,096	0.7	1.000	1.000
UDN	http://udnnews.com/FLASH/	Stock and Financial	1,127	0.7	**0.760**	1.000
CTimes	http://news.chinatimes.com.tw/news/papers/online/	Society	643	0.4	1.000	1.000
CTS	http://www.cts.com.tw/news/headlines/	International	1,064	0.5	1.000	0.959
TTimes	http://www.ttimes.com.tw/	City	1,966	0.7	0.997	**0.530**

TTimes was closed at February 21, 2001.

We choose several Chinese news sites from the data sets. Since news articles of different categories may be published with different presentation styles, we choose one category from each site as shown in Table 5. That is, each site's category indicates a page set used in InfoDiscoverer. For efficient execution, we do not run InfoDiscoverer for all instances in the page set since some sites may contain thousands of pages. Instead, ten training pages are randomly selected from each cluster in the first experiment.

---

8 We check the top-20 ranked pages and manually assign their properties, TOC or article.

To evaluate the performance of InfoDiscoverer, we estimate recall and precision rates of extracted features based on features in the answer set. Regarding features extracted from hand-coding informative content blocks as desired features (the answer set), measures of recall and precision are shown in Figure 16. Clearly, The ideal is that both recall and precision rates equal ones.

Recall rate = Common features / Desired features

Precision rate = Common features / Discovered features.

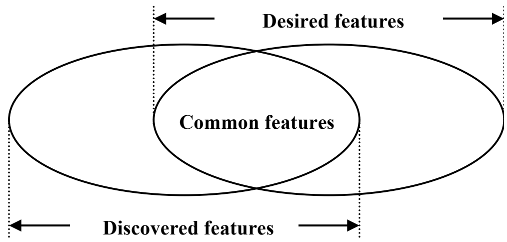


Figure 16: Recall and precision rates of content block evaluation

Based on the greedy heuristic defined in Section 4, InfoDiscoverer dynamically finds the optimal threshold of each site, collects features of extracted informative content blocks, and calculates the corresponding recall and precision rates based on the answer set. The result is shown in the last three columns of Table 5. The result shows that all sites, except for UDN, have very high recall rates (at least 0.956). These optimal thresholds of sites are distributed from 0.1 to 0.7. For example, CNet is converged at 0.5 with recall 0.956. That is, optimal thresholds vary among different sites. The recall of UDN (0.760) is not perfect. By tracing back to the training pages and the corresponding hand-coding data, we found that the hand-coding data of UDN is incorrectly classified because of the inclusion of the title information of news categories. The precision of TTIMES is 0.530 at the optimal threshold 0.7. We checked news articles of TTIMES and found that each page includes an extra content block consisting of "anchors of related news", which are news pages related to the current article. Since the block consists of too many anchors, the text length of the block is even longer than that of the article in many pages. Therefore, these included noisy features affect the decision of the threshold. Consequently, InfoDiscoverer is able to dynamically find the optimal threshold of content block entropy for different page sets.

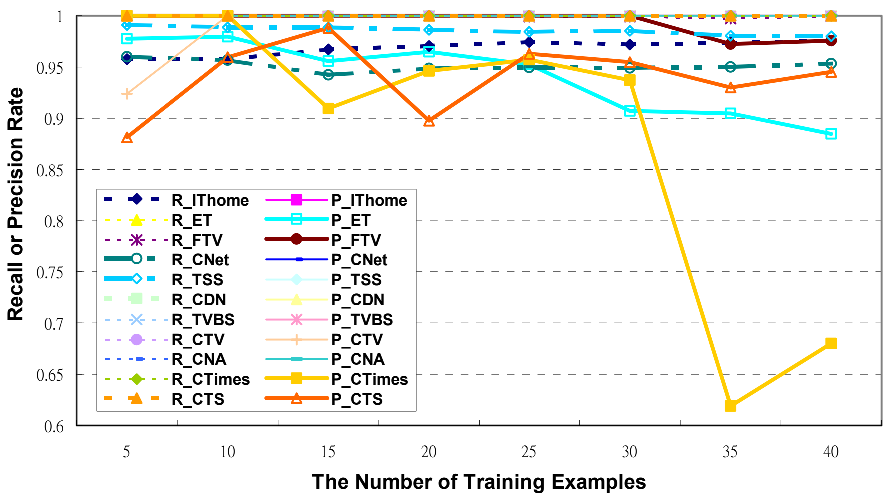


Figure 17: Recall rate of each site.

To investigate the effect of the number of randomly selected training examples, we redo the same experiments on all page clusters. Since UDN has wrong hand-coding data and pages of TTiems contain semantically ambiguous content blocks of related news, both sites are not included in the experiments. The number of training examples starts from 5 to 40 with interval 5. The result is shown in Figure 17, in which the dotted line denotes the recall rate (R) and the solid line represents the precision (P). Most clusters have perfect recall and precision rates approaching to 1 (many R or P lines are overlapped at the highest value 1.0), but precision rates of few clusters (solid lines) are not when the number of randomly selected examples is increased. It is noted that the number of example has an influence on the precision rate since the precision rates of CTS, ET, and CTimes are degraded below 0.9 when the number is increased. In contrast, the random number has little effect on the recall rate since most dotted lines have recall rates larger than 0.956, except for CNet's 0.942. Intuitively, if contents of a cluster are similar, the more examples involved, the higher entropy-threshold would be selected for filtering informative content blocks. Consequently, more training examples do not imply higher precision. However, the recall rate is not affected because a higher threshold means more features included.

6. Conclusions and Future Work

In the paper, we propose a system, composed of LAMIS and InfoDiscoverer, to mine informative structures and contents from Web sites. Given an entrance URL of a Web site, our system is able to crawl the site, parse its pages, analyze entropies of features, links and contents, and mine the informative structures and contents of the site. With a fully automatic flow, the system is useful to serve as a preprocessor of search engines and Web miners (information extraction systems). The system can also be applied to various Web applications with its capability of reducing a complex Web site structure to a concise one with informative contents.

During performing experiments of LAMIS, we found the HITS-related algorithms are not good enough to be applied in mining the informative structures, even the link entropy is considered. Therefore, we developed and investigated several techniques to enhance the mechanism proposed. We conducted several experiments and showed that LAMIS-LN-CB-HR-TW was able to achieve the optimal solution in most cases. The R-Precision 0.82 indicates the enhanced LAMIS performs very well in mining the informative structure of a Web site. The result of InfoDiscoverer also shows that both recall and precision rates are larger than 0.956, which is very close to the hand-coding result. In the future, we are interested in the further enhancement of our system. For example, the concept of generalization/specialization can be applied to find the optimal granularity of blocks to be utilized in LAMIS and InfoDiscoverer. Also, our proposed mechanisms are significant for and are worth of further deployment in several Web domain-specific studies, including those for Web miners and intelligent agents. These are matters of future research.

References

[1]B. Amento, L. Terveen, and W. Hill. Does "Authority" Mean Quality? Predicting Expert Quality Ratings of Web Documents. Proc. of the 23th ACM SIGIR Conf. on Research and Development in Information Retrieval, 2000.

[2]R. Baeza-Yates, B. Ribeiro-Neto. Modern Information Retrieval. Addision Wesley, 1999.

[3]K. Bharat, M. R. Henzinger. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. Proc. of the 21th ACM SIGIR Conf. on Research and Development in Information Retrieval, 1998.

[4]K. Bharat, A. Broder. Mirror and mirror and on the Web: A study of host pairs with replicated content. Proc. of the 8th International World Wide Web Conference, Toronto, Canada, May 1999.

[5]K. Bharat, A. Broder, J. Dean, M. R. Henzinger: A Comparison of Techniques to Find Mirrored Hosts on the WWW. IEEE Data Engineering Bulletin 23(4): 21-26 (2000).

[6]A. Borodin, G. O. Roberts, J. S. Rosenthal, P. Tsaparas. Finding Authorities and Hubs from Link Structures on the World Wide Web. Proc. of 10th World Wide Web Conference, 2001.

[7]S. Brin, L. Page. The anatomy of a large-scale hypertextual Web search engine. Proc. of the 7th World Wide Web Conference, 1998.

[8]A. Broder, S. Glassman, M. Manasse, G. Zweig. Syntactic Clustering of the Web. Proc. of the 6th World Wide Web Conference, 1997.

[9]A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener. Graph structure in the Web. Proc. of the 9th World Wide Web Conference, 2000.

[10] C. Cardie. Empirical Methods in Information Extraction. AI Magazine, 18(4):5-79, 1997.

[11] S. Chakrabarti, M. Joshi, V. Tawde. Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks. Proc. of 24th ACM SIGIR Conf. on Research and Development in Information Retrieval, 2001.

[12] S. Chakrabarti. Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation and Information Extraction. Proc. of the 10th World Wide Web Conference, 2001.

[13] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, J. M. Kleinberg. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. Proc. of the 7th World Wide Web Conference, 1998.

[14] S. Chakrabarti, B. Dom, S. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. M. Kleinberg. Mining the Web's link structure. IEEE Computer, 32(8), pages 60-67, August 1999.

[15] B. Chidlovskii. Wrapper Generation by k-Reversible Grammar Induction. Workshop on Machine Learning for Information Extraction, August, 2000.

[16] C. H. Chang, S. C. Lui. IEPAD: Information Extraction Based on Pattern Discovery. Proc. of the 10th World Wide Web Conference, 2001.

[17] M.-S. Chen, J.-S. Park, and P. S. Yu. Efficient Data Mining for Path Traversal Patterns. IEEE Transactions on Knowledge and Data Engineering, 10(2): 209--221, April 1998.

[18] L. F. Chien. PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval. Proceedings of the ACM SIGIR Conf. on Research and Development in Information Retrieval, 1997.

[19] V. Crescenzi, G. Mecca, and P. Merialdo. RoadRunner: towards automatic data extraction from large Web sites. Proc. of 27th International Conference on Very Large Data Bases, 2001.

[20] B. D. Davison. Recognizing Nepotistic Links on the Web. Proc. of AAAI 2000.

[21] D. Freitag. Machine Learning for Information Extraction. Ph.D. Dissertation of Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, 1998.

[22] 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.

[23] N. Jushmerick. Learning to remove Internet advertisements. Proc. of the 3rd International Conf. On Autonomous Agents, 1999.

[24] H. Y. Kao, S. H. Lin, J. M. Ho, and M. S. Chen. Entropy-Based Link Analysis for Mining Web Informative Structures. The Eleventh International Conference on Information and Knowledge Management (CIKM'02), 2002.

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

[26] N. Kushmerick, D. Weld, and R. Doorenbos. Wrapper Induction for Information Extraction. In Proc. of the 15th International Joint Conference on Artificial Intelligence (IJCAI), 1997.

[27] R. Lempel, S. Moran. The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC effect. In 9th International World Wide Web Conference, Amsterdam, Netherlands, May 2000.

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

[29] S. H. Lin and J. M. Ho. Discovering Informative Content Blocks from Web Documents. The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'02), 2002.

[30] J. C. Miller, G. Rae, F. Schaefer. Modifications of Kleinberg's HITS Algorithm Using Matrix Exponentiation and Web Log Records. Proc. of 24th ACM SIGIR Conf. on Research and Development in Information Retrieval, 2001.

[31] I. Muslea, S. Minton, C. Knoblock. A Hierarchical Approach to Wrapper Induction. Proc. of the 3rd International Conference on Autonomous Agents (Agents'99), 1999.

[32] P. Pirolli, J. Pitkow, R. Rao. Silk from a sow's ear: Extracting usable structures from the Web. Proc. of ACM SIGCHI Conference on Human Factors in Computing, 1996.

[33] G. Salton. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley. 1989.

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

[35] K. Wang, H. Liu. Discovering Structural Association of Semistructured Data, IEEE Transactions on Knowledge and Data Engineering, 12(3):353-371, 2000.

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