Link Structure Graphs for Representing and Analyzing Web Sites

Eduarda Mendes Rodrigues
Natasa Milic-Frayling
Martin Hicks
Gavin Smyth

26 June, 2006

Technical Report
MSR-TR-2006-94

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Link Structure Graphs for Representing and Analyzing Web Sites

ABSTRACT

Standard Web graph representation fails to capture topic association and functional groupings of links and their occurrence across pages in the site.

That limits its applicability and usefulness. In this paper we introduce a novel method for representing hypertext organization of Web sites in the form of Link Structure Graphs (LSGs). The LSG captures both the organization of links at the page level and the overall hyperlink structure of the collection of pages. It comprises vertices that correspond to link blocks of several types and edges that describe reuse of such blocks across pages.

Identification of link blocks is approximated by the analysis of the HTML Document Object Model (DOM). Further differentiation of blocks into types is based on the recurrence of block elements across pages. The method gives rise to a compact representation of all the hyperlinks on the site and enables novel analysis of the site organization. Our approach is supported by the findings of an exploratory user study that reveals how the hyperlink structure is generally perceived by the users. We apply the algorithm to a sample of Web sites and discuss their link structure properties.

Furthermore, we demonstrate that selective crawling strategies can be applied to generate key elements of the LSG incrementally. This further broadens the scope of LSG applicability.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods, indexing methods.*

General Terms

Algorithms, Design, Human Factors.

Keywords

Site structure, hyperlinks, link blocks, Web graph.

1. INTRODUCTION

The Web is a distributed repository of information consisting of billions of inter-connected pages. The link structure within and between Web sites provides the means for browsing through the content and accessing related information on the Web. The content is identifiable by the Universal Resource Locator (URL) that comprises the domain name (e.g., msn.com) and the host that stores and provides access to the particular content (e.g., sport news on sport.msn.com). A URL of an individual page is designated by the full path of the location where the page document is stored in a notional directory structure.

At the same time, link on the pages provide logical organizations of the content that may be significantly different from the content storage structure. When accessing content on the Web users rely on both the URL and the hyperlink structure of the collection of pages. The two structures have been subject of Web research.

Much of the recent efforts focus on the properties of the Web graph that results from hyperlink connections among pages. These properties have been exploited to improve the quality of search engine results [9,12], classify Web sites [1,4], and devise effective compression algorithms for storing the graph [15,17]. However, the link structure of individual sites has not been studied to the same extent. At the same time the studies have shown that users have problems orientating themselves within sites and that current representation of the site, in terms of site maps, for example, have proven to be ineffective [14].

The need for presenting Web pages on small form devices have given rise to methods for analyzing Web page layout to identify coherent content segment and types of links [10, 16]. Link blocks, in particular, have been used to refine page indexing and improve Web search [3]. However, to our knowledge, no attempts have been made to date to unify the page level analysis of links with the commonly used Web graph representation that captures hyperlink relationship among the Web pages. Our work fills this gap and makes novel contributions to the structure analysis of Web content, in particular the Web sites.

We introduce Link Structure Graphs (LSGs) that represent a complete hyperlink structure of the collection of pages, using information about link blocks within pages and the occurrence of blocks across pages. For the sake of concreteness, we apply the method to pages residing on the same host. The method is informed by an exploratory user study that lead to basic LSG concepts, the ‘structural' and ‘content’ link blocks.

We demonstrate that LSG representation can be used to analyze organization of Web sites and illustrate how some of the properties relate to the feedback about sites we received from the users. Finally, we show how LSGs can be generated iteratively by selective crawling. Thus, applications that use LSG may apply such a technique to obtain an approximate LSG structure in cases where the complete information about the site is not available.

In the following sections we provide motivation and objectives of our investigations, describe in detail the algorithm for creating LSGs, and discuss the LSGs for the sample of Web sites that we used in the user study. We contrast our approach with the previous research and conclude with the discussion of selected application areas that will benefit from the LSG representation.

2. MOTIVATION AND PROBLEM STATEMENT

The typical Web graph structure uses distinct Web pages as graph vertices and assigns directed edges to pairs of vertices that are connected by a hyperlink. This approach ignores the associations among the links that are encoded into the HTML representation of the page and that reflect the author's design of the content structure. Such are, for example groups of links that represent navigation menus, or lists of links that collectively cover a particular content, e.g., a list of related news titles, or a table of contents for an online hypertext publication.

Web graph representations tend to eliminate types of hyperlinks that are not deemed relevant for a particular application. Such decisions have an impact on the ability to exploit site structure information in other scenarios. Finally, for efficiency reasons, analysis of the sites is often based on the structure of directories in which pages are stored. This however does not necessarily relate to the logical organization of the content and the users browsing experience.

For all these reasons, we aim at providing alternative representations of the Web and Web site structures that can open up new possibilities in research and applications. We start by investigating basic concepts and ideas that users and Web authors associate with the Web content and page collections.

2.1 Exploratory User Study of Web Page Understanding

Much of the related research on Web design [13] and Web page analysis have introduced concepts that refer to organizations of Web pages and types of links [2,10,16,18]. Such are the concepts of menus, headers, footers, sidebars, and content sections. While this research often includes user assisted evaluation of algorithms, e.g., for analyzing page layout, no study has been conducted to identify how people perceive Web pages and links. The objective of our exploratory study is to discover notions that users have about Web sites, organizations of pages, and functions of hyperlinks. In particular, we investigate three aspects:

1. Do Web users perceive and understand different types of links?
2. Are Web users able to detect associations between links present on a page?
3. Do Web users consider some links to be more important than others?

2.1.1 Study design

For the study we used a sample of 21 sites (see following section) and recruited 14 participants, 9 males and 5 females. All participants confirmed that they regularly use the Internet, with an average reported web usage of 25 hours per week.

The study was designed to include three sessions:

Session 1. Participants freely navigated three websites for a brief period of approximately 5 minutes. Based on their impressions of each site, they answered questions concerning the site organization and the importance of links shown on the page. They were also asked to estimate the approximate size of the site.

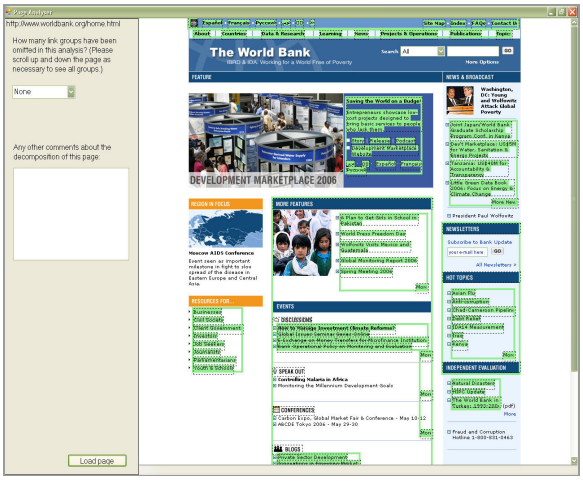


Figure 1. Page Analyzer application that supported the third part of the user study. The application highlights link blocks identified by the LSG algorithm and includes a form for user feedback.

Session 2. Participants were shown two printed pages for each of three sites and asked to consider the links on the pages. They were asked whether the links can be grouped, for example, by content, functionality, etc. Participants were encouraged to freely discuss their impressions of each page.

Session 3. Participants were shown identical pages as in session 2 but this time via a computer-based page analyzer program which detected links and link blocks on each page (see Figure 1). The participants were asked to respond to a set of questions to obtain their opinion whether each set of links formed a coherent group or a menu, and whether there were links on the page that had not been detected by the program. The participants were also asked to rank the prominence of links on the page and provide additional comments as necessary.

2.1.2 Sites Used in the Study

The site sample used in the study comprises 21 Web sites from 7 top-level topic categories of the ODP directory [7] (see Table 1). The sites were divided into seven groups, each containing three Web sites (see Table 2). These sites have also been used elsewhere for analysis of Web evolution [11]. Each participant viewed content from three sites derived from one of the groups.

2.1.3 Findings of the Study

The participants' feedback, derived from the questionnaires in sessions 1 and 2, provides insights into users' assumptions and impressions regarding the structure of the Web pages and the groupings of links on the page. For the sake of space, we cannot include the full study report but we discuss the relevant findings from individual sessions and the correlation of findings among the sessions.

Table 1. DMOZ topic categories from which the sample sites were selected.

1. Arts	**2.** Computers	**3.** Health	**4.** News	**5.** Reference	**6.** Science	**7.** Society
A. Directory **B.** Literature **C.** Television	**A.** Internet **B.** Software, Graphics	**A.** Conditions, Diseases **B.** Occupational and Safety	**A.** Newspapers **B.** Weather	**A.** Libraries **B.** Education	**A.** Institution **B.** Math **C.** Earth Sciences	**A.** Issues **B.** Government **C.** Law

Table 2. Groups of sites for individual participants, distributed to cover multiple topics and size ranges.

Group	**Site**	**Topic**	**Size**		
			Small <10K	**Medium 10-100K**	**Large >100K**
1	1. eserver.org	1.A		✓	
	2. www.hopkins-aids.edu	3.A	✓		
	3. www.acsh.org	7.A	✓		
2	4. etext.lib.virginia.edu	1.B			✓
	5. www.cancerbacup.org.uk	3.A	✓		
	6. www.worldbank.org	7.B		✓	
3	7. www.pbs.org	1.C			✓
	8. www.osha.gov	1.B			✓
	9. www.irs.gov	7.C		✓	
4	10. www.sigmaxi.org	6.A	✓		
	11. elib.cs.berkeley.edu	5.A		✓	
	12. www.artifice.com	2.B	✓		
5	13. www.biostat.wisc.edu	6.B	✓		
	14. www.berkeley.edu	5.B		✓	
	15. www.boston.com	4.A			✓
6	16. www.usgs.gov	6.C		✓	
	17. www.wdvl.com	2.A		✓	
	18. www.wunderground.com	4.B			✓
7	19. nnlm.gov	5.A		✓	
	20. www.eff.org	2.A		✓	
	21. www.nws.noaa.gov	4.B			✓

Session 1 observations

During session 1, participants rated 24 sites out of 42 visited sites as providing well organized content. The rating was based on their initial impressions and limited browsing. When asked to indicate and rate the importance of links on pages, they considered 30 out of the total 42 visited pages as having links of varying importance in terms of content and navigation. The analysis of the participants' written elaborations revealed that their opinions were primarily influenced by the page layout, for example, the presentation of information, location of links on the page, presence or absence of sidebars, screen clutter, and similar.

Finally, in order to validate the participants' estimations of the size ranges for the sites, the Web size distributions for all 21 sites was first calculated from the index sizes of the sites provided by three search engines (Google, MSN and Yahoo!) – see Figure 2. Table 2 shows the size range of the sites. 23 out of 42 user's estimations of site sizes agreed with the estimates we have.

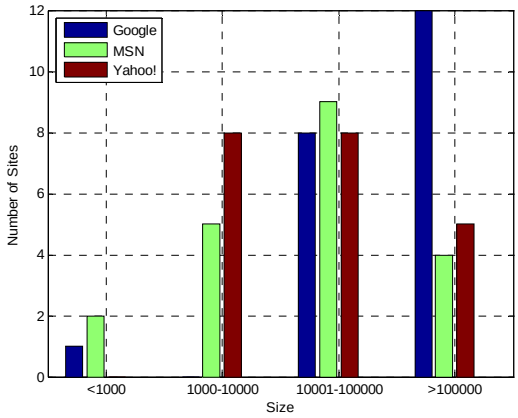


Figure 2. Size range distribution for our sample of sites.

Session 2 observations

In order to explore further how users perceive links or whether they associate several links on Web pages, we prompted each participant to consider printed Web pages and point out, in their own words, the types of links on the pages (2 pages per each of 3 sites). The findings revealed that participants characterized links as: 1) content or topic links relating to the content on the site, 2) navigation links for navigating away to other parts of the site or to external sites, 3) administrative links referring to company information, privacy policy, sitemaps, and similar, or 4) used other attributes to describe the links, such as general purpose links, 'housekeeping' links, internal/external links, etc.. Figure 3 shows participants' categorization of links.

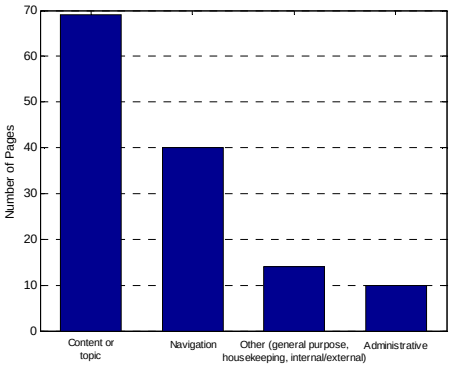


Figure 3. Characterization of links on the Web pages by participants.

Although all the participants were found to group links on the page, they were individually influenced by the layout and presentation of the specific material presented to them. In certain cases participants grouped links by associating them to headings positioned above the links on the page. More importantly, when they attempted to relate blocks of links across two distinct pages from the same site, some of the participants were confused by the difference in appearance and presentation of information.

They stated, there were unsure whether the pages were from the same site. This finding raises obvious implications for Web page design.

In addition, we observed individual differences detecting or categorizing links. Two participants independently revealed that they ignored some content presented on the right of the page. One of the participants elaborated: 'I expect the most important links to be shown on the left as I naturally read from left to right'. When categorizing links, different participants used different terminology to describe essentially the same types of links. For example, links presented at the bottom of the page such as company information, privacy policy, site maps, and similar, were independently described in the function of 'administration', 'bureaucracy', or 'footnotes' to the page. Interestingly, one of the participants discussed links exclusively in terms of internal or external links, whereas another characterized the links only as navigation related or not.

Correlation of results across sessions

Cross session analysis revealed some correlation and discrepancy in participants' comments across sessions. In several cases we can observe the partial consistency of judgments regarding the importance of links. One participant explicitly noted the importance of links on a page from one site during sessions 1 and 2 but did not rate them as prominent when answering questions about automatically discovered link blocks using the on-line page analyzer in session 3. However, another participant referred to the importance of 'content' and 'navigation' links for two sites during session 2 and confirmed this statement in the prominence ratings using the page analyzer.

Finally, five participants explicitly referred to certain links being more important on the page than others. The logs of the page analyzer from session 3 also confirmed that the participants had rated these links as prominent or very prominent, and in some cases, assigned the same attributes to the groups of links, differentiating between prominent and non-prominent groups of links on the page.

Summary

From the comprehensive data we collected in different sessions we outline the most informative findings for the purpose of our research:

- *Perceived importance of links.* The users could articulate and elaborate on different functions and importance of links on the pages.
- *Structure of the page.* The layout, organization of content, and location of links influence the user's perception of the site usability and functions of the links.
- *Structure of links.* Participants could outline the groupings of links and refer to their functions. In some instances they assigned the importance rating to the groups of links.
- *Link categories and difference in terminology.* The user referred to some types of links using different terminology. However, there is commonality in the types of links identified, such as content and navigation links (see Figure).
- *Web site size estimates.* Although only briefly exposed to the site content, the users were able to provide estimates of the site size and based them on various visual and organizational aspects of the pages they observed.

2.2 Basic Concepts and Objectives of the Link Structure Analysis

From the study we outline several generic concepts that we expect to be relevant across Web sites. They enable us to introduce semantics for describing site structure and devise a parsing technique that would be applicable to most of the sites. First, the links appear to be grouped by various types of associations. Second, among link groups we can single out those with a particular function, i.e., the links that enable navigation or high level organization of the site. The pages may have 'content areas' that contain groups of links or individual links that provide access to another HTML page or to any other type of content, such as such as PDF, PS, DOC files, and similar.

Based on this, we set our objectives to define operational definitions of the three types of concepts that we use in the site structure representations:

- *Organizational and navigational link blocks.* These are blocks of links typically repeated across pages with the same layout and underpinning the organization of the site. They are often link lists uninterrupted by other content elements such as text. For simplicity we refer to them as *structural link blocks*.
- *The content link blocks.* are expected to be grouped by content association. However, they are not likely to be repeated across pages but rather point to information resources.
- *Isolated links.* Besides the link blocks, pages often contain links that are not part of a link group and may be only loosely related to each other, for example, by their location within the same paragraph of text.

With regards to the structure analysis algorithm, our objective is to achieve properties that current representations do not have:

- *Locality.* The algorithm should enable us to identify both the global site structure and the local structure around individual pages. Pages that are deemed important by the site author are usually accessible via navigation menus or a site map. However, a large proportion of pages are not accessible from the main navigational structure. In order to help the users orientate themselves it is important to identify relevant site context for each page. We expect these to be sub-structures, i.e., sub-sites within sites. These are often visually recognized from the page templates or navigation elements.
- *Completeness.* The algorithm should capture and represent the complete hyperlink structure of the site. In particular we

aim to represent both navigational structures and logical structures that are explicitly created by Web authors to organize the content. For example, a shopping site author may choose to include a structural element called “best sellers” on a particular page, which links to a set of popular product pages. Although these links are referential in nature, their purpose is still to organize content into a logical structure.

- *Scalability*. The algorithm should run efficiently for arbitrarily large Web sites. Furthermore, the expectation is that exploiting reusability of link blocks across pages will yield alternative graph structures that are of order of magnitude less than the traditional page-hyperlink representations of the Web site structure.

3. Link Structure Graph (LSG)

In this section we define the basic graph concepts and describe the LSG graphs.

3.1 Definitions and Notation

Typically, the Web link structure is represented as a directed graph $G=(V,E)$ where a vertex or a node $p \in V(G)$ represents a Web page and an edge $e=\{p,q\} \in E(G)$, represents a hyperlink from page p to page q.

Generally, in directed graphs we use the notation $p \rightarrow q$ to indicate that p is joined to q by an edge. The vertices p and q are said to be *adjacent* if there is an edge in E that connects them. The *order* of the graph is the cardinality of its vertex set, i.e., $|V(G)|$, and the graph *size* corresponds to the cardinality of its edge set, i.e. $|E(G)|$.

The *in-degree* of a vertex p, denoted as $d^-(p)$, is the number of edges in E that connect other vertices in V to p,. Such vertices represent the *in-neighbourhood* of p and is denoted as $N^-(p)$. The *out-degree* of p, denoted as $d^+(p)$, is the number of edges in E that connect p to other vertices in V, and its *out-neighbourhood* is denoted as $N^+(p)$.

In contrast to the traditional Web link structure, the nodes in the LSG are link blocks. Analogously to the page in- and out-neighbourhood, we define:

- the *link block in-neighbourhood*, as the collection of pages that contain the block of links. In other words, such pages contain links to all the pages whose URLs are in the link block. We also refer to them as *container pages*.
- the *link block out-neighbourhood*, as a collection of all the pages pointed to by the block, i.e., the pages whose URLs are in the link block. These pages we often referred to as *target pages*.

The link between two nodes in the LSG is established when some pages in the out-neighbourhood of one link block are in the in-neighbourhood of the other block. In other words, some target pages of the first link block contain the second block links. The edges of the LSG graph are weighted to preserve information about the aggregated in- and out-degrees of the target pages.

More precisely, if we denote the *target pages* of link block g as $P_g \subseteq V(G)$ and the *target pages* of link block h as $P_h \subseteq V(G)$, then there is an edge between g and h iff $\exists\, p \in P_g$: $\forall q \in P_h$, $\{p,q\} \in E(G)$. Alternatively, if we denote the *container pages* of h as $Q_h \subseteq V(G)$, then there is an edge between g and h iff $|P_g \cap Q_h|>0$.

Note, this definition of link block relationship does not imply that the two blocks are on the same page but rather that by following links from one block one can arrive at a page that contains the links that are in the second block.

3.1.1 Definition of graph node types

A *structural link block* (Section 2.2) therefore comprises a set of links whose target pages contain the block itself. In other words, the target pages are a subset of the container pages for the block. In terms of the Web page graph, the target pages P of the structural link block form a clique, i.e., a subset of vertices that are pairwise connected (see Figure 4 above).

Note that cliques of the site graph may also result from the links that are present within the textual parts of the page. Our algorithm for structural link blocks is conservative in the sense that it does not allow text elements in the block. We denote the structural link blocks as *s*-nodes.

In some instances, incomplete structural blocks may occur, for when the link to the currently viewed page is disabled or dropped from the menu link set. In this case, the link blocks found within the target pages are not identical but have a high degree of overlap. By considering the undirected sub-graph that comprises the set of all target pages we note that it still forms a clique. We use this information to merge the distinct blocks into a unique block that represents the union of the link sets.

A block of content links which are of referential nature tend not to be repeated across the target Web pages. These link blocks are indicated as *c*-nodes. All other links on the page, referred to as isolated links, are for the purpose of LSG representations collected into a single bag-of-links and denoted as an *i*-node.

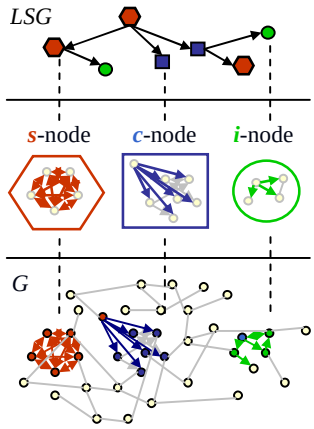


Figure 4. Relationship between LSG graph nodes and edges of the Web site hyperlink graph *G*.

3.2 LSG Algorithm

For a given site, we create the LSG in two phases. In the first phase we analyze layout of individual pages by parsing the corresponding HTML DOM structure in order to identify candidate link blocks. In the second phase we classify the link

blocks into *s*-nodes and *c*-nodes based on the properties of their in- and out-neighborhoods, i.e., their sets of target and container pages. We create links among the blocks to reflect the block containment relationship and include the *i*-nodes into the structure. In the following section we describe in detail each of these phases.

3.2.1 Page layout analysis and link block indexing

The process by which link blocks are detected involves generating the DOM tree representation of the HTML page and parsing each element of the tree using a depth-first search strategy. We are not generally concerned about the complete page segmentation into blocks of content and links.

Link blocks are identified as contiguous hyperlinks lists comprising more than *l* hyperlinks (in our implementation *l*=2 is used) that share the same common ancestor on the DOM tree tag structure. URLs are extracted from the `href` attribute of the `<a>` or `<area>` tags and converted into canonical form. Text formatting tags, such as `<font>` or `<b>`, are not considered for finding the common ancestor tag, as they do not impact on the layout of the links on the page. The algorithm allows for text elements containing non-alphanumeric characters to be located on the same DOM sub-tree that contains the block of links. In other words, white spaces and other characters used often to separate menu items, like '|', are allowed to be in between the list of links. The only text that is preserved from each page is the anchor text associated with each hyperlink, or in the case of image hyperlinks, the `alt` text, if present.

Every time the algorithm comes across a block that has not been previously detected, it stores the block along with the associated DOM path and the unique ID of the container page URL. If a block is repeated across multiple pages, the algorithm only updates the in-neighbourhood of the block accordingly. Links to external Web sites are acknowledged by the LSG representation, by adding the container page to the in-neighbourhood of a global *i*-node that targets external pages.

3.2.2 Classification of link blocks and graph generation

The algorithm proceeds by scanning through all the extracted link blocks and calculating for each pair of blocks the following two parameters, the *resemblance* and *containment* of the link blocks, in case they contain at least one common link. This intermediary step is carried to provide simplification and more granularity of the LSG representation, when possible. The algorithm:

1. Detects the presence of identical blocks with different DOM paths or blocks that are slight variations of each other. For example, the same menu can be present in the top and the bottom section of a given page or there may be some slight differences in the HTML that impact on the extracted DOM paths. Such link blocks can be merged, as they represent the same structure element.
2. Detects the presence of blocks with the same DOM path whose elements are subsets or supersets of other blocks. For example, menus that expand when one of the items is selected are identified as distinct structures. The block that contains both menu and sub-menu items is, in fact, a sub-structure of the block that only contains the main menu items and can be split in two.

Block resemblance is defined as the ratio of links the blocks have in common. Given two link blocks, g and h that target a set of pages P_g and P_h, respectively, their resemblance score R_{gh} is calculated as (1). The containment of block g in block h, denoted as C_{gh}, is characterized by the percentage of links in g that are common to both blocks (2).

$$R_{gh} = \frac{|P_g \cap P_h|}{|P_g \cup P_h|} \quad (1) \qquad C_{gh} = \frac{|P_g \cap P_h|}{|P_g|} \quad (2)$$

If R_{gh}=1, the block with the smallest in-neighbourhood, say h, is removed from the LSG representation, and the in-neighbourhood of the g is adjusted accordingly, i.e. the set of container pages of g is updated: $Q_g = Q_g \cup Q_h$. This provides reduction of the number of LSG nodes, while preserving the page connectivity information.

If C_{gh}=1 and $|P_g \cap Q_h|$=1, i.e. h is a superset of g and there is only one target page of g that contains block h, the algorithm updates the target set of pages of h as $P_h = P_h \setminus P_g$, and the container set of pages if g as $Q_g = Q_g \cup Q_h$. This procedure, while maintaining the number of LSG nodes, enables to capture more granular link structures.

If both $C_{gh} > \tau$ and $C_{hg} > \tau$, where τ is some predefined threshold for the direct and inverse containment of g and h, the algorithm considers both blocks to be the same link structure. In our implementation, we require that at least 60% of the links of g to be shared with h and vice-versa (i.e. τ=0.6). This heuristic acknowledges that some types of navigation menus, either by design or inadvertently, are missing links. Thus, it relaxes the first condition of total resemblance. Similarly, the algorithm removes the block with the smallest in-neighbourhood, say h, from the LSG, and the in- and out-neighbourhoods of the g are adjusted accordingly, i.e. the set of container pages of g is updated, $Q_g = Q_g \cup Q_h$, as well as the set of target pages, $P_g = P_g \cup P_h$.

The LSG representation preserves aggregated information about in- and out-degrees of individual pages, by associating weights with the graph edges. The weight of an edge between two LSG nodes g and h, equals the total number of in-links to the overall set of target pages of the block P_g originating from target pages of link block h, P_h.

The next step of the algorithm is to classify each unique link block as either *s*-nodes, *c*-nodes or *i*-nodes. The isolated links, extracted as bag-of-links from the pages, are classified as *i*-nodes without further processing. To distinguish between *s*-nodes and *c*-nodes we look at the reusability of the link block across its target pages.

If a block g is only contained in a single page, but all target URLs are internal page references, the algorithm classifies the block as a *s*-node. Otherwise, it classifies it as a *c*-node. If g is contained in multiple pages, the algorithm checks if the block targets form a maximal clique of the (undirected) site graph G, by analyzing the overlap between container Q_g and target P_g pages of g. If the overlap is above a certain threshold, $|Q_g \cap P_g| > \tau$ (again, we use τ=0.6), the link block is classified as a *s*-node. Otherwise, it is classified as a *c*-node. Navigation menus with hyperlinks to pages from a different host within the same domain would fall under this last category of *c*-node.

4. ANALYSIS AND GENERATION OF LSGs

We have applied the LSG algorithm to the same sites from the user study and we have carry out site structure analysis based on the resulting graphs. We have also investigated how the LSGs can be generated following a selective crawling strategy. This section reports on the experiments and the results that were obtained.

4.1 Analysis of the Sample Sites

4.1.1 Characterization of the sample sites

In Section 2, we briefly described the sites that were presented to the users, mentioning the estimated size of each of them based on the index sizes of three search engines. Our crawl of the same sites visited only HTML content, as opposed to other type of content on the Web, and retrieved pages residing on the same host as the homepage. Thus, the number of pages crawled was in many cases lower than the sizes reported by the search engines. Figure 5 shows the differences between search index sizes and our crawl. The possible incompleteness of our crawls actually highlights one major benefits of the LSG representation: our graph model has been formulated such that LSGs can be generated incrementally, given any sample of pages from the site.

Further properties about the organization of the Web pages within the directory structure and the link structure of the sites, in terms of minimum distance from home page, are shown in Table 3 and Figure 6. It can be observed that despite the differences in size, most pages of the sites are usually 3 to 5 clicks away from the home page and that the directory level where most pages reside is typically close to the root directory (1 to 3 levels down the directory tree).

4.1.2 LSG properties analysis

The LSG algorithm was applied to our crawls and structure graphs were generated in each case. An example LSG is shown in Figure 7, and was obtained for site number 6 (www.worldbank.org), whose crawl was around 11 thousand pages. The *s*-node only LSG reveals the compactness of this particular Web site navigational structure. These structure *s*-nodes touch around 20% of the pages on this host and about 68% of the pages contain at least a navigational link block. The statistics for the remaining sites are listed in Table 4.

The relative spread and range of influence of *s*-nodes and *c*-nodes is evident from Figure 8. From the graph it is apparent that some sites, have a wide spread of structural link blocks and are likely to have a large number of pages sharing the same template, while others are very rich in content link blocks. For e.g. about 80% of pages from site 21 can be reached from link blocks within other pages. It is apparent that structure properties are heterogeneous across sites of the same size range.

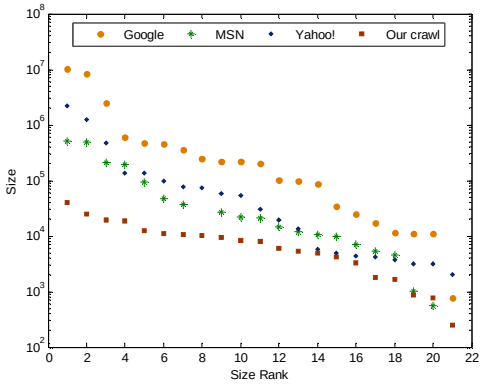
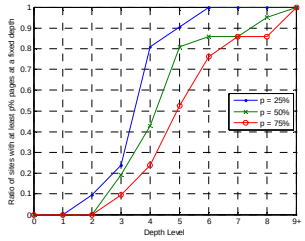


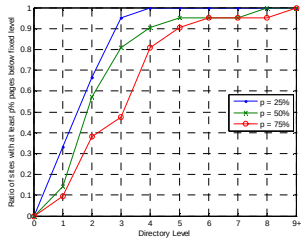
Figure 5. Estimated sizes of the sites as indicated by Google, MSN and Yahoo! indexes, and the actual size of our crawl. The sites are sorted by size rank order.

Table3. Minimum distance from homepage (depth level) and directory level where the highest percentage of pages is located.

Site	Depth level	Ratio pages	Dir. Level	Ratio pages
1	3	0.53	1	0.36
2	4	0.37	2	0.60
3	9+	0.50	2	0.98
4	4	0.49	1	0.90
5	5	0.29	6	0.32
6	9+	0.52	2	0.45
7	5	0.38	4	0.40
8	4	0.44	2	0.76
9	5	0.32	2	0.69
10	2	0.37	2	0.87
11	9+	0.40	9+	0.49
12	3	0.29	1	0.36
13	5	0.29	3	0.24
14	4	0.47	4	0.53
15	5	0.45	3	0.46
16	5	0.28	1	0.86
17	5	0.35	3	0.38
18	3	0.73	2	0.43
19	5	0.58	1	0.54
20	3	0.40	2	0.31
21	6	0.53	3	0.71



a)



b)

Figure 6. Ratio of sites having a percentage of pages *p* (*p*=25%, 50%, 75%) at any given **a)** depth or **b)** directory level.

We have also analyzed the connected components of the LSGs to gain further insight into the navigation properties of the sites. We considered only *s*-nodes and the edges connecting them to verify what portion of the site a user would potentially be able to access using only navigation menus on the pages starting their navigation from one of the *s*-node container pages. Disconnected *s*-node components may be an indication of the presence of multiple templates or sub-sites within the main site.

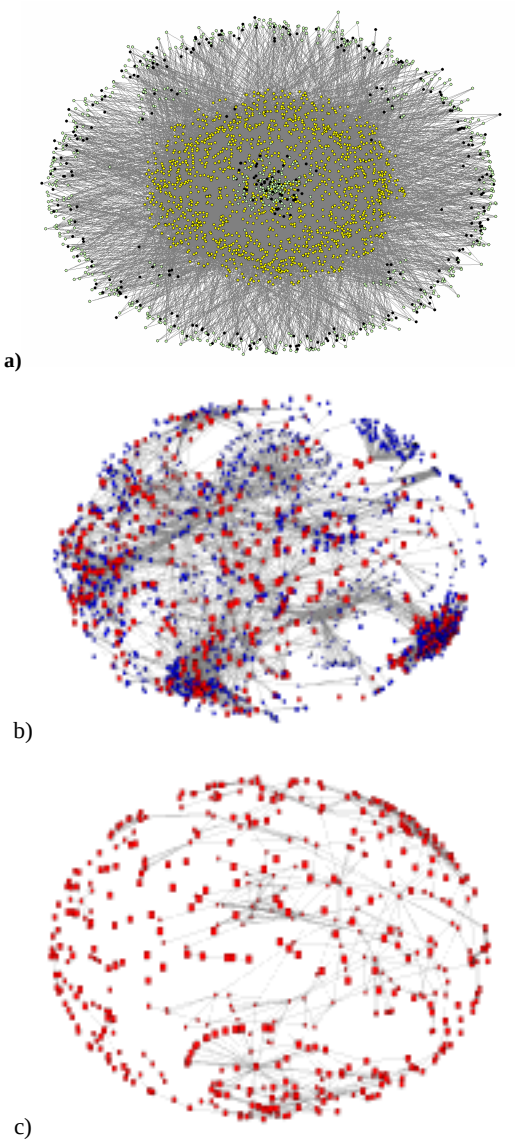


Figure 7. **a)** Standard Web site graph and LSG obtained for www.worlbank.org, showing **b)** *s*-nodes and *c*-nodes **c)** *s*-nodes only.

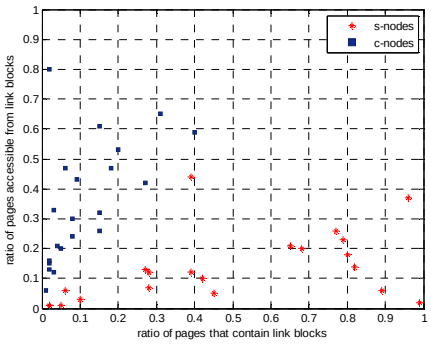


Figure 8. Scatter plot for all sites of the ratio of target vs. container pages of *s*-node and *c*-node link blocks.

Table4. Range of influence and spread of *s*-nodes and *c*-nodes.

Site	Ratio of target pages of		Ratio of container pages of	
	***s*-nodes**	***c*-nodes**	***s*-nodes**	***c*-nodes**
1	0.06	0.61	0.06	0.15
2	0.21	0.32	0.65	0.15
3	0.02	0.06	0.99	0.01
4	0.01	0.15	0.02	0.02
5	0.07	0.16	0.28	0.02
6	0.20	0.30	0.68	0.08
7	0.13	0.43	0.27	0.09
8	0.10	0.65	0.42	0.31
9	0.18	0.42	0.80	0.27
10	0.26	0.47	0.77	0.06
11	0.01	0.53	0.02	0.20
12	0.12	0.47	0.39	0.18
13	0.05	0.21	0.45	0.04
14	0.23	0.20	0.79	0.05
15	0.06	0.13	0.89	0.02
16	0.37	0.26	0.96	0.15
17	0.14	0.59	0.82	0.40
18	0.01	0.33	0.05	0.03
19	0.44	0.12	0.39	0.03
20	0.12	0.24	0.28	0.08
21	0.03	0.80	0.10	0.02

We examined the number of strongly connected components (SCCs) present in each LSG, the ratio of *s*-nodes out of the total number of *s*-nodes and the overall ratio of pages target by the SSCs. We also examined the strongly connected components of the undirected *s*-node LSGs. The aim of this analysis was to identify loose connections between sub-structures of the Web sites.

From Figure 9 it can be observed that the largest SCC of the directed structure graph generally includes a small percentage of the total number of *s*-nodes, while for most sites, the largest SCC of the undirected graph includes a very large percentage of the *s*-nodes. Figure 10 shows that generally sites contain many small SCCs and few large SCCs.

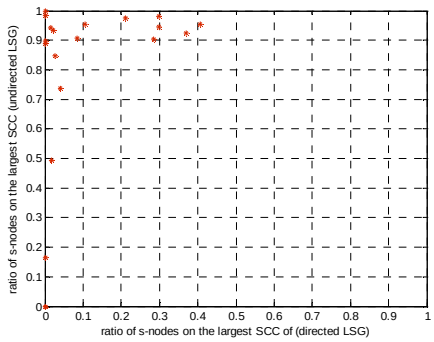


Figure 9. Scatter plot for all sites of relative size of the largest strongly connected component of the LSGs, considering directed and undirected edges.

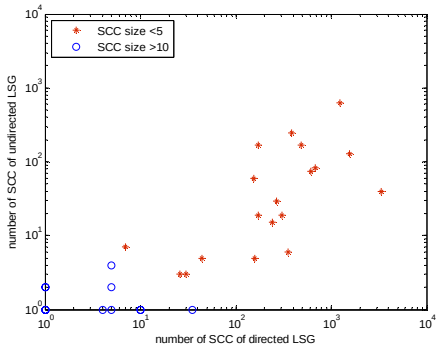


Figure 10. Scatter plot for all sites of the number of small SCCs (size<5) and larger SCCs (size>10)

4.1.3 Correlation with the user study findings

It is interesting to point out that some of the users comments about the sites have some correlation with the structure properties. For example:

Site 4 comments: "It is not obvious how I can get at the content I want by hierarchically navigating menus". "I think this is a website where you need to know what you are looking for. There is a lot of work in reading the text to find out how you need to make the next step in finding what you want".

Structure properties: very few pages contain and are targeted by *s*-node and *c*-node link blocks.

Site 6 comments: "I got very confused in which part I was. The breadcrumbs said I was in 'about us' and I was looking at project information. I think I would use the search bar to get the information that I want."

Structure properties: the site has many *s*-node disconnected components. There are only 10 SCCs with more than 5 *s*-nodes each and those components only touch 4% of the pages.

Site 9 comments: "Easy to navigate the required information, again through the use of grouping and in this case dual menu structure which effects the links available within the side bar which is useful". "The high-level topics and search bar are always available. More specific subtopics can be navigated with a panel that changes, suitably to the context"

Structure properties: 80% of pages contain *s*-nodes and about 20% of pages of the site are accessible through *s*-nodes. There are 17 SCCs with more than 5 *s*-nodes each and collectively touching around 5% of the pages (which is over 1300 pages) of the site.

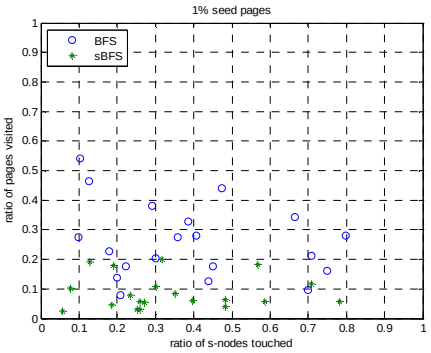
Site 3 comments: "Somewhat overwhelmed by the amount of menus and navigational possibilities. Not clear what the right hand column's task is The top menu is clear. The left menu is context sensitive but not always clear what is happening". "The main topics are clearly shown on the top frame and the left and right frames help as well. Everything seems to be in the right place."

Structure properties: s-nodes target about 2% of the pages but are present in about 99% of the pages. The SCC of the undirected LSG includes all the *s*-nodes.

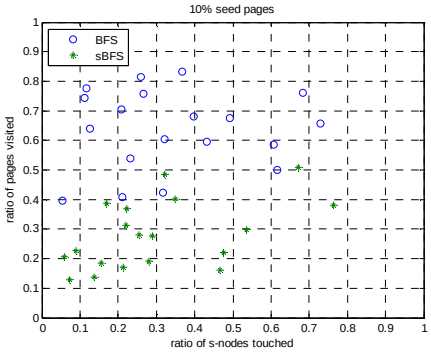
4.2 LSG Generation Through Selective Crawling

In this section we analyze the generation of the LSGs by seeding a crawler with a small percentage of pages from each Web site. The way the algorithm was formulated enables it to create the structure incrementally as new pages are visited. To classify the link blocks as either *s*-nodes or *c*-nodes a crawler needs only to do a 1-step breadth-first search (BFS) of the link blocks out-neighbourhoods and check for the reusability of the blocks across the target pages. Keeping this in mind, we seeded a crawler with 1% and 10% of pages from the sites and we analyzed the number of *s*-nodes that could be identified with 1-step BFS as well as the number of pages that would be visited. We also analyzed the structure generation with selective crawling of hyperlinks which were only part of link blocks (sBSF). The seeding was done by randomly selecting pages from the site.

The results of the experiments are shown in Figure 11. It can be observed that for some sites a high percentage of *s*-nodes can be identified with only 1% of seeds. It can also be observed that in general the selective crawler is able to generate a relatively large part of the LSG graph while visiting significantly fewer pages. Furthermore, it can be observed that having more seeds does not reflect on a much better structure recovery (Figure 11.b). This analysis also points out that the LSG creation does not depend on the order in which the pages are accessed and that the graph can be incrementally expanded as new pages are visited.



a)



b)

Figure 11. Scatter plot for all sites of the ratio of pages visited vs. the number link blocks classified as s-nodes, using as seed **a)** 1% and **b)** 10% of pages

5. RELATED WORK

Throughout the paper we referred to the work and ideas that are related to our effort. Here we elaborate further on related research that is either relevant to the concepts and algorithms we introduced or are potential areas of applications of our method that we intend to cover.

Several approaches have been previously explored to identify content-coherent page segments on the pages, including identification of geometric properties of the page from the HTML page layout [7], vision based segmentation [3,10] and decomposition based on the HTML DOM tree [2,5,12,13]. The later approach has proven effective for detecting and distinguishing template and content segments in Web pages. We adopt very simple heuristics to arrive at efficient segmentation and use the cross page information to confirm the extent of the link block. Our objective is not a full page segmentation but a generic page link block analysis that can be further refined for a particular application.

Block-level page ranking algorithms (PageRank and HITS) have been proposed by [3]. The blocks are used as information units for calculating ranking scores for pages and improving Web information retrieval. The blocks are not restricted to links only, as ours. They are more inclusive of content while our method aims at characterizing the link structure.

The idea of a two-layer graph model of Web link structure has been explored in [15]. They aggregate pages at the domain, host, or directory levels, and create a supernode. Each supernode consists of a hierarchy of pages dictated by the URL hierarchy. This approach is appropriate for modeling and analyzing large Web collections where intra and inter links among pages can be ignored.

Several approaches have been taken to extract logical domains of Web collections through identification of entry pages and exploiting information about the directory hierarchy that contain these pages. Entry pages are determined based on heuristics. Other important pages of the domain are then identified through the in-links property and content relevancy. The boundaries of the domains are established by setting a threshold for the total importance of pages within the domain. Multi-resolution sitemaps are generated. [15]. We expect that the use of LSGs will yield alternative methods for decomposition of sites into domains.

In relation to identifying the structural elements of the sites, the authors in [17] present an algorithm that extracts the skeleton of web sites, i.e., the structure used to organize content. The algorithm identifies navigation links from pages. However, unless sub-domains are connected from the homepage of the site through navigation links, their structure cannot be represented.

6. DISCUSSION AND FUTURE WORK

With increased use of the Web as a platform for building applications and services, it is essential to provide methods for efficient characterization of the Web organization Our work has focused on the representation of link properties of the Web pages that reside on the same host, The user study has shown that such collection of pages are recognized as a unit and given organizational attributes.

However the LSG approach is applicable to any sub-structure of the Web. The distinction between external and internal links is used only to scope the page collection to be described. In the current approach the external links are used in the representation either as parts of the link blocks or a collection of external isolated links. The only difference in treatment is that external links are not examined further for reusability of link blocks.

The LSG representation complements the traditional graph representations of Web structures, having the following advantages. It provides:

- Compact representation of the complete hyperlink structure.
- Integrated representation of the fine level analysis of link associations, based on the HTML DOM structure, and the reuse of links across pages, as designed by the author
- Ability to generate incrementally the main organizational and content elements of the LSG structure, i.e., the s-nodes and c-nodes.

We illustrate how the LSG representation can be used to analyze navigational properties of the site. However, we should also point out the unique and essential aspect of LSG. The core elements of the LSG, i.e., *s*-*c*-*i*-nodes, relate to the visual organization of the page structure. The blocks of links found in the page layout structure. This makes the use of LSG particularly promising as the foundation for creating models of the Web site structure and exposing such structure to the user. Indeed, one of the challenges in building effective support for navigation is in connecting the abstraction of the structure, e.g., a graph, with the 'page level' experience of the user while browsing and searching the content.

Considering the computational aspects, in order to create an LSG structure for a site we incur a cost above the simple Web site graph. The parsing of the HMTL structure to identify link blocks does not represent a significant overhead. In order to maintain the generality of the method we use simple heuristics for identifying link blocks and the parsing of the HTML is efficient. We incur a higher cost in the second pass that involves categorizing link blocks that were found from the HMTL DOM analysis.

We envision several types of applications of LSG representations. First, we expect that the link properties between the *s*-nodes and *c*-nodes will enable us to identify the logical sub-structures of the site. These sub-structures could be used for various purposes, including the presentation of search result within the site structure and revealing the overview of the site structure that is sensitive to the user's current navigation path. As we have demonstrated, the LSG structure can be created incrementally as the user browses the collection of pages.

Second, we anticipate that the LSG can serve as the basis for tools that web designers, authors, and administrators may be able to use, controlling the navigation properties and optimizing the reusability of links and content across the pages.

Finally, characteristics of the link blocks, and inter-block relations in LSGs can augment the search algorithms and further improve the performance [16].

Each of the applications may require further refinement of the algorithms for generating the appropriate LSG structure. Current approach is designed for generality and illustration of the basic principle. We expect that variations of the LSG structures, optimal for a particular application scenario can be devised. However, it is

essential to gain good insights about the usage and that calls for a user informed approach to the algorithm design as we illustrated through our exploratory user study.

7. REFERENCES

[1] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer, "The Connectivity Sonar: detecting site functionality by structural patterns," In: *Proc. of the* 14th ACM *Conference on Hypertext and Hypermedia* (HT), Nottingham, United Kingdom, August 2003.

[2] Z. Bar-Yossef and S. Rajagopalan, "Template detection via data mining and its applications," In: *Proc. of the* 11*th international conference on World Wide Web*, WWW'02, pp 580-591, 2002.

[3] D. Cai, X. He, J.R. Wen and W.Y. Ma, "Block-level link analysis," In: *Proc. of the* 27th *Annual International* ACM *SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pp. 440-447, 2004.

[4] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks. In: *Proc. of the* ACM SIGMOD *Int. Conf. on Management of Data*, pp. 307–318, June 1998.

[5] B.D. Davidson, "Recognizing Nepotistic Links on the Web," AAAI-2000 *Workshop On Artificial Intelligence For Web Search*, pp. 23-28, Austin, Texas, July 2000.

[6] S. Debnath, P. Mitra and C. Lee Giles, "Automatic extraction of informative blocks from webpages," In: *Proc. of the* 2005 ACM *Symposium on Applied Computing*, SAC'2005, pp.1722-1726, 2005.

[7] DMOZ – Open Directory Project: http://dmoz.org

[8] D. Gibson, K. Punera and A, Tomkins, "The volume and evolution of web page templates," In: *Proc. of the* 14*th international conference on World Wide Web*, WWW'05, pp 830-839, 2005.

[9] J. Kleinberg, "Authoritative sources in a hyperlinked environment," In: *Proc. of the Ninth Annual* ACM-SIAM *Symposium on Discrete Algorithms*, pp. 668-677, 1998.

[10] N. Milic-Frayling and R. Sommerer, "SmartView: enhanced document viewer for mobile devices," *Microsoft Research Technical Report* MSR-TR-2002-114, November 2002.

[11] A. Ntoulas, J. Cho and C. Olston, "What's New on the Web? The Evolution of the Web from a Search Engine Perspective," In: *Proc. of* WWW'2004, New York, USA, May 2004.

[12] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank citation ranking: bringing order to the Web," *Technical report, Stanford Digital Library Technologies Project*, 1998.

[13] Jakob Nielsen and Marie Tahir. Homepage Usability: 50 Websites Deconstructed. New Riders Publishing, 2002.

[14] Site Map Usability: 28 design guidelines based on usability studies with people using site maps. Nielsen Norman Group Report, 2002.

[15] S. Raghavan and H. Garcia-Molina, "Representing Web graphs", In: *Proc. of the* IEEE *Intl. Conference on Data Engineering*, ICDE'03, pp. 405-416, March 2003.

[16] R. Song, H. Liu, J.R. Wen and W.Y. Ma, "Learning important models for web page blocks based on layout and content analysis," In: ACM SIGKDD Explorations Newsletter, 6(2), pp. 14-23, 2004.

[17] T. Suel and J. Yuan, "Compressing the graph structure of the Web," In: *Proc. of the Data Compression Conference*, pp. 213-222, 2001.

[18] L.Yi, B.Liu and X. Li, "Eliminating noisy information in Web pages for data mining," In: *Proc. of the Ninth* ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining*, SIGKDD'03, pp. 296-305, 2003.