Topology–based discovery of navigation structure within websites[*]

Filip VOLAVKA, Martin SAJAL, Vojtěch SVÁTEK

Department of Information and Knowledge Engineering,
University of Economics, Prague
Nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
svatek@vse.cz

Abstract. Link topology analysis at the level of website can contribute to the discovery of semantics of individual pages. Most information related to the semantics can be derived from the position of page wrt. the so–called navigation structure of the site. A simple algorithm for navigation structure discovery has been designed, implemented and tested on real data.

Keywords: WWW, link topology.

1 Introduction

Most research in the so–called *webgraph* (i.e. link topology) analysis is currently oriented on large–scale phenomena, such as overall Internet connectivity patterns or virtual web communities, cf. [1]. Purely quantitative computational methods, such as calculation of entropy measures or propagation of weights in a network (e.g. in hub–authority analysis) are generally applied. However, link topology analysis also makes sense in small–scale: that of *websites* belonging to individual organisations. Quantitative methods are useful even in website context, provided we have no significant prior knowledge about the site structure [3, 5]. If, however, we limit our analysis to a *particular class of sites*, we can take into account more specific assumptions about the semantics of topology patterns. Such 'prior knowledge' is particularly useful if the sites are small and quantitative regularities thus unreliable.

In our *Rainbow* project [6] (see also http://rainbow.vse.cz), we consider topology as one of data types for website analysis, aiming at automated creation of *semantic annotations*. Topology analysis is typically much faster than other methods, since it deals with substantially smaller data (hyperlinks represented as source–target pairs); those have often been collected in advance, in the site download phase. However, it offers limited material for semantic interpretation unless combined with other types of analysis, such as that of URL strings, free text, HTML code or metadata. For example, for *company sites*, the role of page as 'hub' in a topology may be positively

[*] The research is partially supported by the grant no.*201/03/1318* of the Grant Agency of the Czech Republic, "Intelligent analysis of the WWW content and structure".

Lubos Popelínský (ed.), DATAKON 2003, Brno, 18.-21. 10. 2003, pp. 1-4.

correlated with its (HTML–based) classification as ‘product catalogue’—the result obtained by more sophisticated methods may thus be confirmed or questioned.

We hypothesise that topology analysis may be beneficial for two interconnected tasks: *semantic classification* of pages, and identification of pages belonging to the same *logical document*. Preliminary analysis of data (cf. section 2) suggested that most information for either task can be derived from the discovery of user–oriented *navigation structure*. The assumption of existence of such structure represents quite widely reusable ‘prior knowledge’, as mentioned above.

2 Preliminary analysis of data

In the initial survey, we collected the topologies of 75 randomly selected *company websites*. By their predominant structure, we distinguished the following types:

- Single file (i.e. page without links, or with outward links only).
- Hierarchy[1] with limited connectivity of pages in the same layer.
- Hierarchy with *complete connectivity* of a set of sibling pages in the second, and possibly subsequent, levels of hierarchy. The union of such sets will be denoted as *navigation structure* (NS), since it enables easy navigation over mutually related topics; the individual sets (possibly multiple ones at the same level, but each with a different parent page) will be called *NS components*.
- Chain (sequence) of pages.
- Sites committed to MM Flash (beyond the reach of our topology analysis).

The results are in Tab.1 (more details in [4]); among the 47 NSs, 38 had their components only in the second layer of the hierarchy, the remaining 9 also in one or more subsequent layers. Since the NS patterns were frequent and conspicuous (could be determined by topology alone), we took their discovery as starting point.

Tab.1. Frequences of topological structures in initial survey

	Abs. freq.	Rel. freq.
Single file	12	16 %
Basic hierarchy	10	13 %
Hierarchy with navigation structure	*47*	*63 %*
Chain	5	7 %
Completely for MM Flash	1	1 %

3 Algorithm for discovery of navigation structure

We only sketch the basic steps of the algorithm. More details can be found in [7]:

1. Collection of *initial set I* of candidate pages. Beginning from a *start–up page*, pages referenced by links are iteratively added. Pages beyond the current server are not considered.

[1] In general, we did not distinguish between hierarchies with one–way and two–way child–to–parent links.

2. *Adjacency matrix* and *minimal–distance matrix* of *I* is constructed. *Depth* of each page is computed as minimal distance from the start–up page.
3. In each set of pages with same depth, *NS components* are subsequently sought, as sets of at least three pair–wise interconnected pages. Their union is output as the resulting navigation structure.

A side–product of the algorithm is the *compactness* measure of the website, computed from the minimal–distance matrix. It roughly expresses the accessibility of information, and thus indirectly the quality of website design.

As a follow–up step, a tentative form of website ‘core’ discovery, as a particular type of *logical document*, was also implemented. The ‘core’ only contains the pages that point to *all* pages from the NS. *Semantic classification* of pages has been designed at informal level but not implemented yet.

The algorithm has been implemented as a collection of PHP scripts, with both *HTML* and *web service interface*. The former also provides webgraph visualisation and displays statistical information (e.g. adjacency and minimal–distance tables).

4 Test results and their evaluation

The first experiment was carried out on six new websites, selected with respect to small size allowing their complete *visual examination*. Tab.2 lists the results; the figures correspond to numbers of pages in the respective sets.

In all cases, a NS *with single component* was found, which always corresponded to some kind of *menu bar* in the browser. Even the *website core discovery* seemed to perform well: the 12 *excluded* pages (11 % of the sample) were mostly (10 cases) relevant for the user *in specific situations* only (login to private parts of the site; pages offering extra services beyond the main occupation of the company; online books/articles). Often, the pages had different outlook, (as presumably based on a different HTML template and created by different people), hence could be understood as *logical documents* of their own. The remaining two cases corresponded to a page *under construction* and to the single ‘error’ of the algorithm: a content page, which (probably by omission of its author) differed from related pages by missing links.

Tab.2. Test results of the algorithm—first experiment

Site no.	Candidate set	Navigation structure	Website ‘core’	Pages excluded from ‘core’
1	64	9	56	8
2	13	6	10	3
3	10	8	9	1
4	7	6	7	0
5	5	4	5	0
6	13	6	13	0

For the second experiment, a larger and more representative sample of data was used: 42 sites of ‘organisations offering products’ (nicknamed as OOP) randomly chosen from the 'Business' branch of *Open Directory*[2]. The size of the sites ranged between 1 and approx. 200 pages, the average value was 52. For 20 of them (47.6%), a NS was identified; most of the other sites, namely 15 (35.7%), consisted of a single page. The NSs consisted of one (most often, always in depth 1) up to 9 components; the average value was approx. 2. The average size of NS component was approx. 6, which corresponds to 'reasonable' size of menu bar.

Unfortunately, the *website core discovery* part of the algorithm did not scale well to larger sites, since unacceptably many useful content pages were excluded.

5 Future work

Since Rainbow primarily aims at semantic analysis, the most important next step is to implement the mapping from the output of the described algorithm to *semantic classes* of pages. The OOP data currently serve for training/testing all analytical modules of the *Rainbow* system. This should enable to match the topological classes of pages (and other web resources) with classes determined by other methods; we already adapted a method originally used for *ontology merging* for this purpose [2]. Statistical correlation among the classes could then be used for *co–operative* analysis including multiple modules. Finally, we should start to take *frames* into account, since the current algorithm does not process them correctly.

References

1. Second Workshop on Algorithms and Models for the Web-Graph (WAW 2003), `http://www.almaden.ibm.com/cs/people/ravi/waw2003.html`.
2. Labský, M., Svátek, V.: Ontology Merging in Context of Web Analysis. In: Workshop on Databases, Texts, Specifications and Objects (DATESO'03), Ostrava 2003.
3. Mathieu, F., Viennot, L.: Local Structure in the Web. In: Poster Session of the International World–Wide Web Conference, Budapest 2003.
4. Sajal, M.: Analysis of link topology among web pages. (In Czech) [Master thesis], University of Economics, Prague, 2002.
5. Skopal, T., Snášel, V., Svátek, V., Krátký, M.: Searching Internet Using Topological Analysis of Web pages. In: WWW Based Communities for Knowledge Presentation, Sharing, Mining and Protection, CIC 2003/PSMP: Las Vegas, USA.
6. Svátek, V., Kosek, J., Labský, M., Bráza, J., Kavalec, M., Vacura, M., Vávra, V., Snášel, V.: Rainbow – Multiway Semantic Semantic Analysis of Websites. Accepted for the DEXA Int'l Workshop on Web Semantics, Praha 2003.
7. Volavka, F.: Website core identification based on link topology. (In Czech) [Master thesis], University of Economics, Prague, 2003.

[2] http://www.opendir.org