Node Classification in Open Hypermedia Systems

Javed Khan
Creative Computing
Coventry University, UK
jak786@hotmail.com

ABSTRACT

From online learning websites to social networks, Hypermedia systems can be found anywhere on the Internet. These systems offer great navigational freedom to the internet users with its rich link structure that interlinks different type of nodes. However, the complexity of this linked structure contributes significantly towards the *Lost in Hyperspace* problem.

In this paper we propose our node classification technique which helps users in identifying different nodes’ type, hence, providing navigational assistance to the users. The technique attempts to classify hypermedia nodes on the basis of their specific properties (i.e. HTML tags) and identifies them as landmarks to represent a simplified view to the user, thus, enhancing the overall navigational experience of the online user.

KEYWORDS

Hypermedia Structure, Usability, Hierarchical Structure, Node Classification.

1. Introduction

“Hypertext” refers to a set of information nodes connected with a sets of links in between the nodes, when the information stored in the nodes are not only text, but also image, audio, animation and video, is known as “Hypermedia”. Hypermedia structure refers to the underlying link structure between the nodes.

The main strength of hypermedia systems is that they have a flexible structure and give user great deal of freedom to browse and interact with the information contained within them [1]. It is precisely this freedom that attributes to many problems; one of the problems that seem inherent within hypermedia systems is that users tend to lose their way in the maze of information within the system. This is commonly referred to as the ‘Lost In Hyperspace’.

Nielsen describes ‘LIH’ as “one of the major usability problems with hypertext” [2] and suggests that to help users navigate the hyperspace; assistance must be provided to understand and recognise their present location in the overall structure. A logical progression would be to apply the techniques, which are used to organise physical space into electronic space. So the users can apply their navigational skills learnt in physical space, to navigate them in electronic environment.

In the domain of Urban Design, Lynch concluded that people represent their physical environment (as mental or cognitive maps), he found that subjects frequently recall particular features, such as nodes (*hubs*), landmarks (*a recognizable point of reference in the larger space*), paths (*routes between locations*) and edges (*boundaries*) and these are used effectively for navigation in physical environment [3]. To capitalise on this, several researchers have tried to experiment, to what extent navigational strategies from the physical environment can be applied to the electronic environment. In particular, Botafogo et al. [4] and Modjeska & Marsh [5] tried to identify landmarks within electronic environment, so that the users can apply their navigational skills learnt in physical environment, to navigate them in electronic environment.

To facilitate navigation in electronic environment the environment must be organised in a structure. The question arises here, which organising structure is the best suited, to structure a knowledge domain? In this paper we argue that hierarchical structure is the best structure to organise the information domain.

The remainder of this paper is structured as follows: Section 2 surveys the major problem faced by hypermedia users such as ‘LIH’. Different hypermedia structures are explored in Section 3. We conduct an experiment to analyse websites in Section 4. In Section 5 we extend the ideas from Section 4 to propose our node classification technique; and finally, Section 6 concludes.

2. Lost In Hyperspace

Often the user lacks an intuitive sense of location and orientation within the hypermedia structure, not knowing where the user currently is, in the overall structure and which is the best way to proceed from there? Computer Science literature refers to this as the ‘LIH’ problem.

Foltz and Davis [6] believed that ‘LIH’ problem is due to a lack of informed design of the space and insufficient navigational affordances, particularly to describe and to locate the physical environment.

The problem appears to be one of unfamiliarity with the hypermedia structure rather than user being unfamiliar with the information content [1].

In an experiment, Elm & Woods [7] gave users of a hypertext system a set of information retrieval tasks and found that the degree of ‘LIH’ experienced by the subjects...
was independent of their level of expertise in the information domain.
As a result of their study [7] three different forms of being lost were outlined:

1. Not knowing where to go next.
2. Knowing where to go, but not knowing how to get there.
3. Not knowing where they are in the overall structure.

‘LIH’ problem appears to be a navigational related problem, to tackle this efficiently; navigational strategies used in physical environment must be extended and adopted to assist navigation in electronic environments.

3. Hypermedia Structures

Many possible structures have been suggested to address the problem of hypermedia system structuring to minimize cognitive overhead and maximize coherence?

Three typical structures used in experiments are “hotwords”, “hierarchical maps”, and “spider maps”. “hotwords” are simply links embedded in the body of the text. “Hierarchical maps” provide a graphical representation of the information in a hierarchical structure similar to an organisational chart. “Spider maps,” also called “concept maps,” provide a more detailed graphical representation of the information with more extensive non-linear cross-linking, similar to a spider web. Findings from Beasley & Waugh [8] concluded that there was a significant difference in perceived disorientation, with subjects in the “hotwords” condition having the most disorientation while subjects in the “hierarchical maps” condition having the least [8].

Another study comparing “hotwords” versus “hypermap” structures revealed no difference in recall but found that the group using “hotwords” reported feeling significantly confused and frustrated [9]. One explanation of why “hotwords” perceived as confusing and frustrating might be because “hotwords” demand additional cognitive resources from the user. For instance, the users have to remember long trails of link, in order to get explanation of unfamiliar words in the knowledge domain.

In another study, Modjeska & Marsh [5] conducted experiments on websites, which represented two structural types (strongly & weakly hierarchical) and two sizes (small & large). Their experiment produced a number of preliminary results. On website scale, the structure had a significant effect on user navigation, with strongly hierarchical sites having a greater number of nodes accessed than did the weakly hierarchical sites. The website structure had significant effects on user perceptions with strong hierarchical sites being perceived as smaller than the weak hierarchical sites. They concluded that strong hierarchical structures appear more usable than the weakly hierarchical ones.

3.1 Discovering Hierarchies in Hypermedia Structure

Several literatures [1,4,5,8] suggests that hierarchical structure is the best-suited structure for making user’s mental model about the system, and then a method to discover hierarchies in the existing hypermedia system is required so that any new node can be found automatically.

Botafogo et al. [4] reports that the authors of hypermedia systems are encouraged to create hierarchical structures, but when writing, the hierarchy is lost because of inclusion of the cross-reference links. They suggested the ways of recovering the lost hierarchies and finding new ones; they also suggested helping authors by identifying the properties of the hypermedia structure by the use of the metrics, these metrics can be used to capture the useful properties of the hypermedia structure.

To find hierarchy in a structure, two tasks must be performed: first the root node must be identified and then hierarchical and cross-referential links must be distinguished, and they recommend following as the fundamental properties of the Root Node:

1. It has to reach almost every node in the hypermedia.
2. The distance from the root to any other node should not be too large.

If the distance from the root to a node is very large, there will be a navigational problem. Users will have to go through a long path before reaching the desired information.

3. Finally the root must have a reasonable number of children.

4. Our Experiment

We conducted an initial experiment to extend Botafogo et al. approach; the purpose of the experiment was to encode the hypermedia structure into a distance matrix. So we can help users in identifying landmarks within the structure and to make additional observations which could aid navigation in the hypermedia structure.

An html parser (written in Java) was used to crawl through links on a web page to capture complete link structure of a website in a distance matrix, example is shown in figure 2 to illustrate the idea.
The experiment was designed to count number of in-links (links from other pages pointed to the current web page) and out-links (links pointed to other web pages). Web pages with the highest number of out-links were classified as index pages, and web pages with the highest number of in-links were classified as reference pages.

The experiment successfully identified the index and reference pages from the captured websites but just by identifying index and reference pages does not provide sufficient help to the user to efficiently navigate the structure, especially in large & complex hypermedia structure.

A basic idea is that with a book, there is a certain set of expectations or schema of what a book is, and how it is comprised. We know that we can look in the index to find out the location of information. As familiarity with the text grows, the reader becomes more familiar with the various landmarks in the text and the relationship between them. The reader is in effect building up a mental model of the text, based on the orientation cues.

These orientation cues, which are so important for constructing a mental model, are absent in many hypermedia systems [1].

In line with the above, we observed from our experiment that more comprehensive node classification is required to identify various different types of nodes. This will make the underlying structure more comprehensible by providing simplified view to the users and may minimise Lost in Hyperspace problem.

5. Node Classification

The classification of the nodes must produce more categories than just index and reference nodes in order to analyse every node’s position in the overall structure and possible usability problems within the hypermedia structure.

We classified nodes into following categories:

**Self-referencing Node:**
These nodes contain links on the top and are broken into parts such that each link on the top points to the different part of the same node. They can be easily identified by a crawler because their link contains # to denote self-referencing.

**Dead End Node:**
These nodes do not have any out-links.

**Index Node:**
This type of nodes contains a large amount of out-links, and their distance to other nodes is relatively small. Also they are one of the only nodes, which can reach most of the nodes in the site. Site Map of a website is a good example of such nodes.

**Entry selection Node:**
This type of node was seen on start of website, either to choose the country of their residence or the language. This node usually points to different parts (sub-structure) of a hypermedia. Their out-links point to the (root of) sub-structure.

**Information Node:**
These nodes have relatively high in-links than out-links. The in-links value might be very low or 0 (dead-end nodes can also be information nodes).

**Applet Node:** These types of nodes contain applet tags and can be categorised as an applet node.

**Data Entry Node:** Nodes, which contains a number of form tags this, would suggest that user could enter data (user response). In addition, these nodes have a less number of out-links i.e. Submit button.

**Search Node:** These types of nodes contain a searching mechanism.

**File Download** – Pages with file downloads like pdf or exe files.

6. Conclusions and Further Work

In this paper we have discussed the major problem faced by the users of hypermedia system such as the “Lost in Hyperspace” (LIH) problem. We have cited previous research which establishes that it’s a navigational related problem. The solution to this problem is to enhance
navigation in hyperspace by employing techniques used for navigation in the physical environment. To provide navigational assistance to the user, the hypermedia system must inform the users about the sense of distance, direction and their current location in the overall structures; similar to a geographical map of a region. Briefly, identifying landmarks in hypermedia structure would provide a simplified view of the structure; it would reduce the cognitive overhead and would assist navigation in the hypermedia structure.

In order to get better understanding of the problem, we conducted an experiment based on Botafogo et al. approach to identify landmarks such as root nodes within hypermedia system.

One of the key finding of our experiment was that, just by identifying root nodes doesn’t really solve LIH problem. We believe that LIH phenomenon can be minimised by identifying nodes into several classes, for instance, self-referencing nodes, data-entry nodes, etc. This might help users to remember, what kind of node they are looking for? And could apply search filter based on the user required node class.

Another of our key finding was that there should be different classes of links, each class can be represented by different colour for feedback to the user and it will make crawling efficient because crawler can be programmed to handle different link class differently.

Furthermore, we believe that once the landmarks have been identified, additional Meta data could be associated with it to list all the important attributes of the landmark such as, Name, No of children, Creation Date, Last modification Date, No of Visits (No of visits can be displayed as heat maps).

These landmark attributes would have to be updated whenever Meta data is updated but if we can automate this process then it would make the crawling a lot more quick because we just would have to read this Meta data of all the landmarks with the hypermedia system and report to the authors on a real-time basis. This will provide very useful information to the hypermedia authors.

Limitation:
Our current crawler can’t read-in secure sockets and the authentication mechanism. Currently, we are not parsing cascading style sheets attached with a webpage.

Advantages:
Current implementation can pick up broken URLs quickly and can make a graphical representation of the website link structure. We can merge landmark attributes data with the website link structure, for instance no of visits to a particular landmark could be represented as heat map on the website map.

Possible application
The crawler can be used to build up the website link map and visual tool (under development) can display this map to the user with user’s current location in the overall structure and possible next moves within the structure.

For future work, we are looking to conduct a real user experiment to demonstrate the effectiveness of our node classification proposed in this paper.

We are working on a tool which will show simplified view of a hypermedia system to authors and users. Our next article will be based on this tool.

7. REFERENCES