Semantic Segmentation of Hypertext on the Basis of Automata Model

Alexander A. Sytnik
Head of the chair of information systems and technologies
Sergey V. Papshev
Associated professor of the chair of information systems and technologies
77, Politehnicheskaya St., Saratov State Technical University, Saratov, Russia, 410054
tel/fax: +7(845 2) 526 660, e-mail: as@sstu.ru, http://www.sstu.ru

Abstract
This article is about website optimization on the basis of semantic segmentation of its hypertext pages. Automata model for hypertext is introduced to describe hypertext organization of website. According to this model states of automaton correspond to hypertext pages and transitions correspond to hypertext links. Then the algorithm of segmentation for hypertext pages based on calculating automata congruences/tolerances is offered. The segmentation is produced by partitioning of states for some congruence or covering of states for some tolerance. The algorithm work under consideration of degrees for hyperlinks and pages, assigned during statistic supervision for website. To achieve appropriative result the initial automaton is consequently modified by decreasing its connectedness.
Keywords: Semantic segmentation; Website; Automata; Congruence; Hypertext.

1 Introduction

Any website, as a rule, provides unidirectional service, which transmit information from a site to some client. In the course of life cycle of a website inevitably some information becomes outdated, becomes not actual. Pages are modified for appeal improvement, new sections of a site are added. Over time the initial hypertext website structure can be broken, and it can bring isolated pages or hyperlinks in anywhere. Therefore the reengineering of a website becomes one of the main tasks of website administration.

During design and support process of large hypertext information systems semantic segmentation is very important to reduce extraction cost of information. Also it would be very useful to make marketing decision about similar pages. It may be also an initial point for constructing of system to anticipate user's requests.

Existing methods of hypertext organization mainly are founded on assumptions about preferences and navigating behaviour of users. Developers often don't formulate these assumptions obviously and don't estimate, whether they correspond to real behaviour of users. It is a serious problem because wrong assumptions about semantic organization can lead to inefficient hyperlink structure.

The modern approach to web analytics of websites is based on statistic analysis of hits for hypertext pages. Main purpose of like analysis is to explore behaviour of cites'

visitors on the basis not only hits, but page viewings (preferably unique viewings). So, it is well known that advertising companies connect the content to web pages in consideration of cookie based information. It make possible to associate such files with one user. Then purpose marking approach was developed as continuing of web-analytic ideas.

The web analytics includes both research of the site client's behavior, and external aspects of its behaviour, such as request reference. This technology allows carrying out classification for sources of request. The page tag technology has become a next step to collect more data about the user.

According this idea complex approach to web analytics was formulated. It means an adaptation of tool, which earlier was adapted for page oriented site organization, to site analytics realizing as tool set of JavaScript scenarios on the basis of Ajax-, Flash-, Java-technologies and Silverlight applications (Clifton, 2008; Kaushik, 2010). In this context the web site consisting of pages, are understood as a set of "places" and "tasks". "Place" is a resource where constantly there are users and tasks arise, when the users have some purpose.

At this point it is actually to present website by automata or graph model to produce website segmentation by partitioning or covering of website pages.

2 Discrete Models for Hypertext System

To formalize the reengineering process of a website it is expedient to construct its mathematical model. In this case it is possible to apply mathematical methods for optimization of its functional structure. As the website consists of a set of information nodes, between which there are connections, to analyze the structure of hypertext is used discrete models. The most obvious model for hypertext system is directed graph. This model allows task's solving in analysis and synthesis of hypertext topology. It make possibility to apply well known analysis and optimization algorithms to the graph model of a website such as calculation of metrics, finding of shortest paths, isolated vertexes, connected subgraphs, etc.

Numerous investigations for optimization of hypertext structures were conducted using graph model (the review of corresponding results is resulted in Hollink, Someren, Wielinga, 2007), however mainly it had deal with only static organization of the hypertext.

It is well known when hypertext space structure for some information domain was formed, it may be used as recognizer for human representation about this object (Duffy, Cunningham, 1998) and wherefore user behaviour track may essentially differ from initial static hypertext tracing. Besides, transitions on hypertext links may be completed browser aides tools. Additional effects may appear during work with hypertext (opening of new windows, script actions and etc.). Finally today many internet websites are designed as software application.

Thus, graph model constructed on the basis static hypertext is limited and can't used to describe the real detailed working process of the user with hypertext information.

Network discrete model of website allows us to investigate not only the structure of the object, but the semantics of its elements, expressed in the weights of vertices and links.

Finite determine automaton may be another functional model for hypertext system. In his paper we will mean automaton as triple $A=(S,X,\delta)$, where $S=\{D_i\}_{i\in I}$ is the set of states corresponding to set of hypertext pages (not only static, but also generated by script tools, and places in pointed above sense), $X=\{w_j, D_i\}_{i\in I, j\in J_i}$ is set of inputs, corresponding to hyperlinks from words $\{w_j\}_{j\in J_i}$ on page D_i, and $\delta: S\times X\to S$ is transition function between pages. This partly defined automaton may be reduced to completely defined automaton by adding loop transitions.

3 Semantic Segmentation of Hypertext

When we use discrete model for website, we may solve some typical tasks for it: find deadlock states, connectivity (strong connectivity) for automata, factorization on some property, etc. Each such task has some substantial interpretation for hypertext information system.

Let's consider more detail the task of semantic segmentation for hypertext pages to optimize hypertext structure of site.

Let $A=(S,X,\delta)$ is finite automaton, which is model of some hypertext information system. The equivalence θ on the set S is named congruence for automaton A, if it is stable concerning transition function δ as it is defined eq. 1

$$(\forall s_1, s_2 \in S)(\forall x\in X)((s_1,s_2)\in\theta \Leftrightarrow (\delta(s_1,x),\delta(s_2,x)\in\theta) \quad (1)$$

I.e. if s_1 and s_2 are in the same θ-class, then states produced under action of any input signal will be also in one class.

Identity relation Δ and universal relation $S\times S$ are congruencies in any automaton.

As an example let's consider the case when in automaton $A=(S,X,\delta)$ there are some states s_1 and s_2 for which for any input signal $x\in X$ $\delta(s_1,x)=\delta(s_2,x)$. Such states are named indistinguishable. It is obvious that the relation of indistinguishability is congruence of automaton A.

The collection of all congruencies of automaton A are denoted as $Con\ A$. It is well known that partly ordering set ($Con\ A$, $\subseteq$) is a lattice. The node "0" in the lattice correspond to identity relation Δ and node "1" correspond to universal relation $S\times S$.

Let $Con\ A$ is a congruence lattice for automaton A. The partition of states of automaton A (pages of site) into equivalent classes is correspond to some congruence $r\in Con\ A$. This partition is produced by congruence in assumptions about paired equivalence of some automata states. We can make these assumptions on the basis of statistic exploration of site page viewings. Partition is defined by congruence allows to structure information of the site, logically uniting automata states, which correspond to

one equivalence class. More detail information about automata algebraic structures may be found in the book of Birkhoff, Bartee (1970).

Analogical procedure may be done if a tolerance relation will be selected instead of equivalence relation. It may be more adequate to some fuzzy classification for site pages. In this case we will have covering corresponding to tolerance relation instead partition corresponding to equivalence relation.

In any case, the resulting segmentation of site pages may be used both to optimize developing technology of site and to market it by escorting viewings of page from the same class by some information corresponding to the class.

3.1 Algorithm of Segmentation

As at congruence and tolerance construction for automaton, the great value has degree of connectivity of automaton, i.e. how many hyperlinks are in website pages. There is a danger that if the website has high degree of connectivity then we can't produce nontrivial congruences and tolerances. To reduce this problem it is necessary to "make more easier" hyperlink structure by deleting some hyperlinks, for example on the basis of statistic information about rarely used hyperlinks.

Let's consider the algorithm of segmentation using congruence construction.

Algorithm *A*. The construction of nontrivial partition of site pages.
Input. Automaton $A_0 = (S, X, \delta)$, which is model of hypertext information system. Weight function $f : \delta(s, x) \to [0..M]$, which set some number from diapason [0..*M*] to hypertext link in consideration of statistics for site visitors.
Output. Set U of partitions for S.
The work of algorithm.

1. Initializing. $i=0$. $j=M$. $U=\varnothing$.
2. On automaton A_i set $Con\ A_i$ is constructing.
3. Nontrivial congruences $Con\ A_i \backslash \{0,1\}$ are added to set U, i.e. $U = U \cup Con\ A_i \backslash \{0,1\}$.
4. If $j=0$ then end (terminate algorithm), else $i=i+1$.
5. On the basis automaton A_{i-1} automaton A_i is constructing by replacing each transition $\delta(s, x) = p$, for which weight function $f(\delta(s, x)) = j$, to transition $\delta(s, x) = s$. Another degrees of transitions are decreased by one.
6. $j=j-1$. Go to step 2.

Each resulted partition of set U may be considered as essential point for semantic segmentation of site.

3.2 Example of Segmentation of Hypertext

Let's illustrate work of algorithm for some example site.

Let some hypertext site consist of four pages. Let discrete model for site is automaton A_0 with set of states $S=\{1,2,3,4\}$, input alphabet $X=\{x_1,x_2\}$ and its functioning is defined transition table (see Table 1).

Table 1: Definition of automaton A_0.

δ	x_1	x_2
1	2	1
2	3	2
3	3	2
4	4	3

The transition diagrams for autonomous component corresponding to inputs x_1 и x_2, presented at the Figure 1. The definition of function f is set by weight cursive numbers at arrows.

The work of algorithm for $A_0 = (S, X, \delta)$ is resulted low.

Step 1. Initializing: $i=0$. $j=2$. $U=\varnothing$.

Step 2. Nontrivial congruences of x_1-component of automaton A_i are set by following partitions (only non one-element blocks are indicated): θ_1=[2,3], θ_2=[3,4], θ_3=[1,2,3], θ_4=[2,3,4], θ_5=[1,2,3,4].

Analogical list for x_2-component $\theta_1, \theta_3, \theta_4, \theta_5$ and θ_2^*=[1,2]. The lattice diagram *Con* A_0 is presented at figure 2. Hence, $U=\{\theta_1, \theta_3, \theta_4\}$.

Step 3. Because $j\neq 0$, then $i=i+1=0+1=1$.

Step 4. On automaton A_0 automaton A_1 is constructed. The transitions diagram new automaton A_1 is presented at the figure 3.

Step 5. $j=j-1=2-2=1$.

Step 6. Nontrivial congruences for automaton A_1 are θ_4 =[2,3,4], θ_6 = [1,2], θ_1 = [2,3], θ_7 = [2,4], θ_3 = [1,2,3], θ_8 = [1,2,4].

Hence $U = U \cup \{\theta_6, \theta_7, \theta_8\} = \{\theta_1, \theta_3, \theta_4, \theta_6, \theta_7, \theta_8\}$.

Step 7. Because $j\neq 0$ then $i=i+1=1+1=2$.

Step 8. On automaton A_1 automaton A_2 is constructed.

The transitions diagram of this automaton consist of only loop transitions with zero weight numbers.

Step 9. $j=j-1=1-1=0$.

Step 10. The lattice *Con* A_2 contains only trivial congruences, then set U is not changed.

Step 11. Because of $j=0$ work of the algorithm is terminated.

Thus we have the set of possible segmentations of four pages:

$U=\{\theta_1, \theta_3, \theta_4, \theta_6, \theta_7, \theta_8\}$.

Let some statistic observation have showed that pages, which correspond to states "1" and "4" were visited from the same website (or visited by the same user), i.e. these states in some sense are equivalent. Comparing this data with results of the algorithm we can get following variants of page segmentation: ([1] [2,3,4]), ([1] [2,4] [3]), ([1,2,4] [3]). If we choose, for example, the third partition then we can predetermine the automaton A_1 to the automaton with outputs so that transitions to states "1", "2" and "4" will be followed by one output signal.

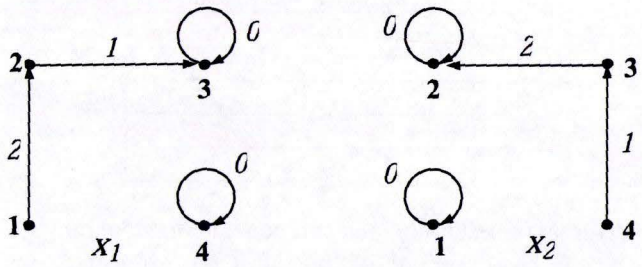

Figure 1: Transition diagram for automaton A_0.

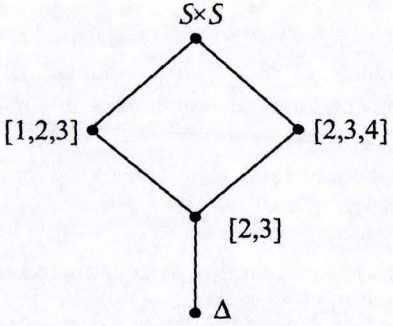

Figure 2: Lattice diagram *Con A*.

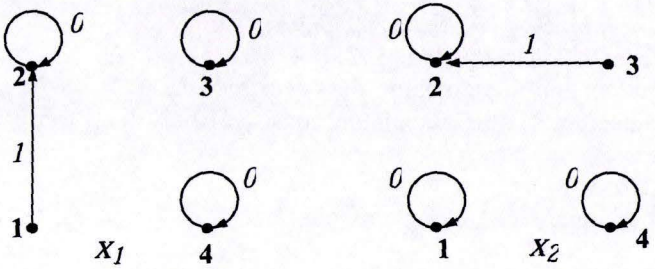

Figure 3: Transition diagram for automaton A_1.

4 Conclusion

Discrete models are used for describing of hypertext systems. Automata model may be successfully applied to describe hypertext system in dynamic using. According to this model states of automaton correspond to hypertext page and transitions correspond to hypertext link. We showed how automata-algebraic methods may be used to solve tasks of semantic segmentation of hypertext information. Then the algorithm of segmentation for hypertext pages based on calculating automata congruences/tolerances

is offered. The segmentation is produced by partition of states for some congruence or covering of states for some tolerance. The algorithm work in consideration of degrees for hyperlinks, assigned during statistic supervision for hypertext website. To achieve appropriative result initial automaton is consequently modified by decreasing its connectedness. The resulted factor automaton constructed on chosen congruence may be used during reengineering process of website in consideration to anticipate behaviour of users.

References

Hollink Vera., Someren Maarten., Wielinga Bob J. (2007) Navigation behavior models for link structure optimization. User Modeling and User-Adapted Interaction. Volume 17, Number 4 / September, pp. 339 - 377.

Duffy, T. M., Cunningham, D. J. (1998) Constructivism: Implications for the Design and Delivery of Instruction. Handbook of Research for Educational Communications and Technology. Edited by Jonassen D. H. New York: Simon & Schuster Macmillan.

Clifton Brian. (2008) Advanced Web Metrics with Google Analytics. Wiley Publishing, Inc., Indianapolis, Indiana.

Kaushik Avinash. (2010) Web Analytics 2.0. The Art of Online Accountability & Science of Customer Centricity. Wiley Publishing, Inc., Indianapolis, Indiana.

Birkhoff Garrett, Bartee Thomas. (1970) Modern Applied Algebra. McGraw-Hill.