Research on the Measuring Model of the Website Structure Optimization Based on the Theory of DIT

Cong Liu
Information Technology Center
Tsinghua University
Beijing, P.R.China
e-mail: lc@cic.tsinghua.edu.cn

***Abstract*—This paper investigates a large number of documents and papers on website structure optimization, proposes a measuring model of the website structure optimization based on the theory of Distance of Information-state Transition (DIT), and makes the model more reasonable and practical by introducing the concept of semantic similarity and Analytic Hierarchy Process (AHP). The case study shows that the model is effective and feasible.**

Keywords-website optimization; DIT; semantic similarity; AHP

I. INTRODUCTION

Compared with webpage information organization, the website information organization is a more advanced form of network information organization. Moreover, the website information organization helps to form a website by organizing the information more ordered and effective with a specific link structure from a higher level. This specific link structure, which is the website structure, is the external structure of webpage. Thus, an important criterion of the judging whether website information organization is effective is whether the website structure is reasonable. However, the design of this kind of structure is considered from website, and is inevitable different from the website users' expectations [1, 2]. In order to reduce the difference, the website structure optimization has been the focus of current research. The achievements can be divided into two categories, which are based on the existing various research, analysis, evaluation and summarization: one is based on the research of the website problems assessment of user behavior data, and another is based on the research of website model without taking account the specific user click behavior.

A quantitative evaluation method for website structure optimization is established in this paper, which is based on the theory of DIT. Its main idea is to quantify the cost of website navigation, which is the standard analysis of evaluating website navigation structure optimization degree. Analysis of calculating the cost data of the website navigation, we can also identify the main factors of the navigation structure optimization degree, allowing the designers or maintainers pay more attention to the improvement of navigation structure.

II. BASIC CONCEPTS

A. The Theory of DIT

The Distance of Information-state Transition (DIT), the ideas and methods of the quantitative measure of information and knowledge, is presented by Professor Wang in 2004[3]. Let the probability of an object transit from state x_i to state x_j be p_{ij}, the measure formula of DIT $DIT(x_i \rightarrow x_j)$ or d_{ij} is (1):

$$DIT(x_i \to x_j) \equiv d_{ij} \equiv \log(1/p_{ij}) = -\log(p_{ij}) \quad (1)$$

where $i,j=1,2,\ldots,n$, $\sum_{i=1}^{n} p_{ij} = 1$.

B. Semantic Similarity

The constraints of the user access requests for the website can be defined as the specific information in the webpage of the website or the semantic descriptions of the services. Usually the specific service theme the website provides is identified and each user request can be mapped to the corresponding service theme. Therefore, the user request can be refined as a service theme corresponding to one or more target columns. The target column is the column page in the website content page the users expect to access.

The semantic relevance can be modeled as the semantic similarity between the service themes and the target columns in the same semantic vector space [4]. Let $S=\{S_1,S_2,\ldots,S_{|S|}\}$ denote all the service themes and $T=\{T_1,T_2,\ldots,T_{|T|}\}$ denote all the target columns in the website, where $|S|$ represents the number of the service themes and $|T|$ represents the number of the target columns. In the same semantic vector space, the semantic vectors of the service theme S_i and the target column T_j are correspondingly defined as $V_{S_i} = \left(A_{S_i,1}, A_{S_i,2}, \cdots, A_{S_i,l}\right)$ and $V_{T_j} = \left(A_{T_j,1}, A_{T_j,2}, \cdots, A_{T_j,l}\right)$ respectively, where $i=1,2,\ldots,|S|$, $j=1,2,\ldots,|T|$, and l is the dimension of the semantic vector space.

Thus, the semantic similarity between the service theme S_i and the target column T_j can be calculated by (2):

$$Sim\left(V_{S_i}, V_{T_j}\right) = \cos\left(\theta_{S_i,T_j}\right) = \frac{\sum_{r=1}^{l} A_{S_i,r} \times A_{T_j,r}}{\sqrt{\sum_{r=1}^{l} A_{S_i,r}^2} \times \sqrt{\sum_{r=1}^{l} A_{T_j,r}^2}} \quad (2)$$

C. Analytic Hierarchy Process

The AHP [5], which is short for Analytic Hierarchy Process, is an effective method breaking a complex problem into several small manageable elements, in order to build a multi-target and multi-level model and generate an ordered and hierarchical structure. The basic idea of AHP to overall determine the multiple element weights of complex issue is to pairwise compare these elements, then to sort these elements by the global weights of them, and finally establish the weight of each element.

III. THE MEASURING MODEL OF THE WEBSITE STRUCTURE OPTIMIZATION

A. Website Structure Description by Graph

The main components of a website are web pages and links, in which the web pages include content pages and index pages (or column page). In most cases, the number of content pages is far more than the number of column pages and each content page at least belongs to a column page. Considering that when users navigate to a content page directly from an upstream column page, it will take relatively less navigation cost. So we mainly discuss how to optimize the website column page in this paper. Thus, the website structure can be abstracted as an un-weighted digraph $G=(N, E)$, where N is the set of nodes representing website columns and E is the set of edges representing links that connect the nodes. Assume that there are C_N columns in the website and their sequence numbers can be marked as $0,1,2,\ldots,C_N$-1.

B. Mathematical Model of the Website Structure Optimization

Evaluating the website structure optimization level can be determined by the cost user access the website information or services paid for (i.e. navigation cost). Website navigation cost reflects the ability to achieve the user visiting purpose under the navigation structure guidance.

Because the DIT provides a quantitative measurement method of information state transition, website navigation cost can be measured by DIT. Shorter DIT means that it is shorter to reach the target state, more convenient to operate, smaller cost to navigate, and vice versa. The default initial state of user access is the website homepage, the target state is the information expected to visit or the target column page of the corresponding service, the process of searching is that the website navigation structure guides the user achieve the purpose of visit.

Let $R=\{R_1,R_2,\ldots,R_{|R|}\}$ denote the user access request set in the website， where $|R|$ is the number of user access requests, and the cost of website navigation can be defined for the sum of DIT all the website user access requests. Such as (3):

$$DIT(R) = \sum_{R_i \in R} DIT(R_i) \tag{3}$$

where $DIT(R_i)$ is the DIT of the user request R_i, $i=1,2,\ldots,|R|$.

IV. CALCULATION METHOD OF WEBSITE STRUCTURE OPTIMIZATION MEASURING MODEL

A. Calculation of the Target Column DIT

1) Finding all the paths of each target column: Referring to the Path Tree Spanning Algorithm [6], find out all the paths of each target column. In order to describe multiple navigation paths to a column page, C_i represents that the number of the paths that can be reached the column i, $S_{i,d}$ represents the number of steps that access the column i in the d^{th} path and the pages in the path is numbered following the order, $M_{i,d,k}$ represents the label of the column k that is in the d^{th} path of the column i, and $P_{i,d,j}$ represents the column page j that is in the d^{th} path of the column i. The path $L_{i,d}$ can be described as $L_{i,d} = \{P_{i,d,1}, P_{i,d,2}, \cdots, P_{i,d,S_{i,d}}\}$, $P_{i,d,1}=0$, $P_{i,d,S_{i,d}} = M_{i,d,S_{i,d}} = i$, where $i=1,2,\ldots,|T|$.

2) Calculation of each path DIT : Information state and its transition probability are the two basic elements of measuring the DIT. So this paper set the rules of information state of each node of paths as the following three sides, and the following calculation of actual example is according to such information state rules.

a) Menu items (menus or links within a page): there are N choices as the forward sequence to the target state, take N to 1.

b) Click: there are two states, i.e. click or don't click, take 2 to 1.

c) Column type: information list choices, take 20 to 1.

The DIT of path $L_{i,d}$ can be calculated as (4):

$$DIT(L_{i,d}) = \sum_{k=1}^{S_{i,d}} [DIT(M_{i,d,k}) + DIT(Click)] + DIT(TYPE_i) \tag{4}$$

where $DIT(M_{i,d,k})$ is the DIT of column $M_{i,d,k}$ which is the DIT of selecting the next column from column $M_{i,d,k}$, usually N is the out-degree of the node, $DIT(Click)$ is the DIT of one click, $DIT(TYPE_i)$ is the DIT of the type of column i.

3) Calculation of each target column DIT: Since there are many paths to each target column, we have to get the path weight to calculate each target column DIT. This paper uses the AHP to construct 5 order judgment matrix (through consistency test) of 5 path types (following the path hops to classify).

The DIT of target column i can be calculated by (5):

$$DIT(i) = \sum_{d=1}^{C_i} \alpha_{i,d} \times DIT(L_{i,d}) \tag{5}$$

$$\alpha_{i,d} = \frac{\varphi(S_{i,d})}{\sum_{j=1}^{C_i} \varphi(S_{i,j})} \tag{6}$$

where $\alpha_{i,d}$ represents the weight of the d^{th} path of target column i, $i \in T$, it can be calculated by (6), $\varphi(S_{i,d})$ represents the corresponding function of the d^{th} path of the target column i and 5 path types index weight.

B. Calculation of Each Service Theme DIT

First, construct the incidence matrix *IM* of the service themes *S* and target columns *T*, and optimize the matrix, $\forall S_i \in S$, $\forall T_j \in T$, if $Sim(V_{S_i}, V_{T_j}) < \eta$, then $Sim(V_{S_i}, V_{T_j}) = 0$, otherwise unchanged, where η represents the threshold of the semantic similarity.

Then, according to *IM*, define the calculation of weight β_{S_i,T_j} of target column $T_j \in T$ in service theme $S_i \in S$ as (7).

$$\beta_{S_i,T_j} = \frac{Sim(V_{S_i}, V_{T_j})}{\sum_{T_q \in T_{S_i}} Sim(V_{S_i}, V_{T_q})} \quad (7)$$

where T_{S_i} represents the target column set in the service theme $S_i \in S$.

At last, the DIT of service theme $S_i \in S$ can be calculated by (8).

$$DIT(S_i) = \gamma_{S_i} \times \sum_{T_j \in T_{S_i}} \beta_{S_i,T_j} \times DIT(T_j) \quad (8)$$

where $DIT(T_j)$ represents the DIT of the target column in service theme S_i, γ_{S_i} represents the cost coefficient of service theme S_i. Ideally, each service theme corresponds to only one target column. However, with the increase of the target column of the same theme, the cost of website navigation will be raised and it will cause problems for users' selection. Therefore we define the cost coefficient of service theme which will make the calculation of service theme DIT and practical operation much more fit. The cost coefficient is set mainly based the target column number in the service theme.

C. Calculation of Website DIT

There are more than one service themes supported in the website, some of which are important, and some of which are unimportant, according to the features of the website. Based on the consideration above, each theme in the set of the service themes *S* was assigned a value to indicate the importance of this theme. The greater is the value, the more important is the theme. Conversely, the smaller is the value, the less important is the theme. Here, the importance of the service theme $S_i \in S$ can be described by ω_{S_i}. Thus, the website navigation cost for all the user tasks supported in the website can be expressed by (9).

$$DIT(R) = \sum_{R_i \in R} DIT(R_i) = |R| \times \sum_{S_i \in S} \omega_{S_i} \times DIT(S_i) \quad (9)$$

V. Case Study

Here is a simple website structure for example, shown in Fig.1. Assume the set of the service themes is denoted as $S=\{S_1,S_2,\ldots,S_{10}\}$, the set of the website columns is denoted as $N=\{0,1,2,\ldots,21\}$ and the set of the target columns is denoted as T={5,6,7,9,10,11,12,14,15,17,18,19,20,21}, where 0 represents the homepage and $|R|$ sets 1000.

According to the user's application habits to visit the website, define the service theme cost coefficient depending on the experiences. Thus in this case study, the data used and generated in the computation of the website navigation cost before and after the website structure optimization are shown in Table 2 and Table 3. Because of the limited space, the incidence matrix between the service themes and target columns is omitted here.

As can be seen from the data shown in Table 1, by adjusting the relations among the website structure, such as delete column 16 and add the paths to reach the target column 19, the DIT of the corresponding target column can be reduced. The local optimization of website structure is achieved at the cost of increasing the DIT of other target columns.

As can be seen from the data shown in Table 2, through improving the semantic similarity between the service themes and the target columns, each service theme involves about 1-2 target columns after adjustment, whereas each service theme involves about 2-4 target columns before adjustment. Not only the DIT of almost every service theme is reduced, but also the improved website navigation cost has been significantly reduced than the one before improvement, which also illustrates that the website navigation structure has been greatly enhanced on the overall optimization degree.

VI. Conclusions

After analyzing the common evaluation methods of the website structure optimization, this paper discusses the measuring model of the website structure optimization based on the novel theory of the DIT. Case study shows that the measuring model is effective and feasible. The model introduced in this paper can not only bring the new ideas and methods to the website structure optimization, but also provide further support for the website designers to develop optimization design of the website.

Acknowledgment

The author thanks Ruiwei Meng for excellent technical support and critically reviewing the manuscript.

References

[1] Czyzowicz J, Kranakis E, Krizanc D, et al. "Enhancing hyperlink structure for improving web performance," Journal of Web Engineering, Vol. 2, pp. 93 – 127, 2003.

[2] R. Srikant, Y. Yang, "Mining web logs to improve website organization," Proceedings of the Tenth International Conference on World Wide Web, Hong Kong, pp. 430 – 437, 2001.

[3] H.C. Wang, DIT and Information, Beijing: Science Press, 2006, pp.23-55.

[4] G.Salton, "Recent trends in automatic information retrieval," In ACM SIGIR Conference(1986). ACM, New York, pp. 1-10, 1986.

[5] J. Fei, Chen M.Y. Chen, "Studying on the leading cadre economic responsibility audit evaluation based on the AHP," Computer Engineering and Applications, Vol. 18, pp. 25-27, 2003. In Chinese.

[6] Y.W. Wang, D.W. Wang, "Multi-objective model for link structure optimization of e-supermarket website," Control Theory & Applications, Vol. 21, pp. 6-10, 2004. In Chinese.

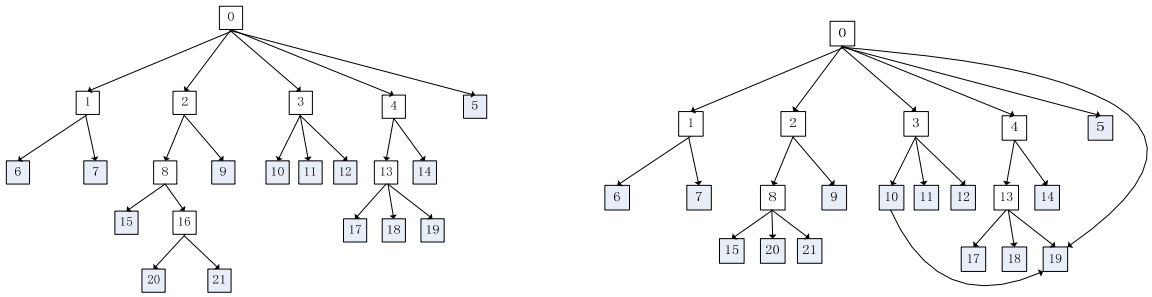

Figure 1. An example of website column structure: the left one is before optimization, the right one is after optimization.

TABLE 1 PATHS AND DIT OF EACH TARGET COLUMN BEFORE AND AFTER OPTIMIZATION

Target column	Before optimization		After optimization	
	Path	***DIT of path***	***Path***	***DIT of path***
5	{0,5}	3.321928095	{0,5}	3.584962501
6	{0,1,6}	5.321928095	{0,1,6}	5.584962501
7	{0,1,7}	5.321928095	{0,1,7}	5.584962501
9	{0,2,9}	9.64385619	{0,2,9}	9.906890596
10	{0,3,10}	10.22881869	{0,3,10}	10.4918531
11	{0,3,11}	10.22881869	{0,3,11}	10.4918531
12	{0,3,12}	10.22881869	{0,3,12}	10.4918531
14	{0,4,14}	5.321928095	{0,4,14}	5.584962501
15	{0,2,8,15}	7.321928095	{0,2,8,15}	8.169925001
17	{0,4,13,17}	12.22881869	{0,4,13,17}	12.4918531
18	{0,4,13,18}	12.22881869	{0,4,13,18}	12.4918531
19	{0,4,13,19}	12.22881869	{0,4,13,19}	8.124037377
			{0,19}	
			{0,3,10,19}	
20	{0,2,8,16,20}	13.64385619	{0,2,8,20}	12.4918531
21	{0,2,8,16,21}	13.64385619	{0,2,8,21}	12.4918531

TABLE 2 PATHS AND DIT OF EACH TARGET COLUMN BEFORE AND AFTER OPTIMIZATION

Service theme	Importance	Before optimization			After optimization		
		Target column	***DIT of service theme***	***Website navigation cost***	***Target column***	***DIT of service theme***	***Website navigation cost***
1	0.1	{5,6,7}	6.632892	15926.2765	{5}	5.584963	10159.5514
2	0.05	{6,7,18}	10.7093		{7,18}	9.71879	
3	0.05	{6,7}	5.854121		{6}	5.584963	
4	0.15	{9,20,21}	18.13245		{9,21}	12.41409	
5	0.1	{10,14,17}	14.22858		{17}	8.124037	
6	0.05	{11,12}	11.2517		{10,11}	11.54104	
7	0.1	{12,15,18}	14.65036		{15}	12.49185	
8	0.05	{14,19}	9.399658		{12,14}	8.456707	
9	0.1	{9,11,20}	17.04398		{20}	12.49185	
10	0.25	{10,17,19,21}	24.36036		{17,19}	10.65237	