User Oriented Link Function Classification

Mingliang Zhu Weiming Hu Ou Wu Xi Li Xiaoqin Zhang

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
Automation Building, No. 95 Zhongguancun East Road
100080, Beijing, China
{mlzhu, wmhu, wuou, lixi, xqzhang}@nlpr.ia.ac.cn

ABSTRACT

Currently most link-related applications treat all links in the same web page to be identical. One link-related application usually requires one certain property of hyperlinks but actually not all links have this property or they have this property on different levels. Based on a study of how human users judge the links, the idea of the link function classification (LFC) is introduced in this paper. The link functions reflect the purpose that links are created by web page designers and the way they are used by viewers. Links in a certain function class imply one certain relationship between the adjacent pages, and thus they can be assumed to have similar properties. An algorithm is proposed to analyze the link functions based on both vision and structure features which simulates the reaction on the links of human users. Current applications can be enhanced by LFC with a more accurate modeling of the web graph. New mining methods can be also developed by making more and stronger assumptions on links within each function class due to the purer property set they share.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data Mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia.

General Terms

Algorithms, Performance, Human Factors.

Keywords

Link function classification, vision features, structure features, web mining.

1. INTRODUCTION

One of the most fundamental differences between web pages and classical documents is that the web pages are hyperlinked to each other, while mining based on hyperlinks has always been an important and effective idea among both webpage-related literatures and real applications.

Currently most link-related applications assume that all links within a certain web page are identical in terms of their utilities and thus the web is modeled as a single-link-typed directed graph. But hyperlinks are not actually the same for human users. Intuitively, page designers put hyperlinks in their webpages for different purposes, while viewers judge and predict each link they see. For example, in the link analysis, all links in a page are assumed to be clicked with the same probability, or imply the same level of recommendation, but actually some of the links are generally more attractive to the viewer. Some applications such as [3] require each link to be a sign of content similarity of the target page, but usually cannot tell which links point to more similar pages without visiting the targets - although this prediction is generally easy for human users. Some methods are proposed to solve these problems. Most of them, such as [4], use clues directly extracted from the DOM structure. However, the underlying DOM tree is usually biased from and much more complicated than the visual representation of the page, especially after the introduction of WYSIWYG html editors. So the structure clues do not always reflect the human designers and viewers. In [1] the vision clues are regarded, but like other methods, it finds a global importance level, or ranking, for links, which is not portable. For example, links of high importance in link analysis do not necessarily leads to high similarity. These motivate us to analyze the functions of links to generate a precise web graph, and try to find the classes and features that match real human users.

2. LINK FUNCTION CLASSIFICATION

The main problem with the single-link-typed web graph is that one application usually requires one certain property of hyperlinks (or one certain relationship between the two linked webpages) but actually not all hyperlinks have this property or they have this property on different levels. Designers set up different hyperlinks for different purposes, which conceal different relationships between the linked pages. On the other hand, it is reasonable that links set up for the same purpose share similar properties. So analyzing the link by functions evidently better fits the real web.

2.1 Major Link Functions

The major link functions are summarized based on study of human behaviors. We examined many webpages to find out for what purposes that hyperlinks are put in webpages. Five main functions for hyperlinks are summarized:

1. *Structural and Navigating*. These links form the hierarchical structure of a website, letting viewers navigate to different parts or topics of the website, or providing some functions over the whole website (such as the “Login” entries).
2. *Indexing and Directory*. They are used as portals to other pages, or providing some function only for the current page. A series of links of this class are often organized together as a list providing a series of choices for viewers.
3. *Citing and Explanation*. They provide extra information about a term mentioned in the page content, such as the explanation of a concept, or details of an event.
4. *Recommending and Expanding*. They provide information in which the viewer of the current page may be also interested.

*This work is partly supported by NSFC (Grant No. 60672040) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453).

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

The target pages may be discussing the same or related subjects, or introducing other valuable stuffs to the viewer.

5. *Advertising and Commercial*. These links are just used for commercial purpose and their target pages usually have little relationship with the containing page.

Compared with the importance level approaches, our function classes are more objectively and impersonally defined so that one can easily tell the function of a link without ambiguity. While the functions of different links have determined, link-related applications can be enhanced with a more precise modeling of the web graph or they can mine on purer web sub-graphs where all links share similar properties.

The *Advertising* class is usually easy to identify by URL based rules, and there are already such methods in use. We now focus only on the first four classes, which are intuitively much more informative but cannot be handled by simple rule based methods.

2.2 Methodology

We use supervised machine learning to train classifiers based on a couple of link features. As a supplement to the common features directly extracted from the DOM structure, we adopt some vision features to describe human reaction of the link. The features we used are as follows:

Vision features: consist of the spatial features and content features. Spatial features describe the hyperlinks' spatial appearance in browsers rather than the position in the DOM tree, including the position of the link and the position and size of the vision block containing the link. The vision blocks are obtained by the VIPS algorithm [2]. The content features include the number of characters in the anchor text and the size of image in the link.

Structure features: provide heuristics on the interaction of the link object and the webpage context in which it lies, including: whether the link is in list or heading; character distances to neighboring links; and character distances to neighboring line-end characters.

Whole page features: heuristics implying the type of the webpage, which may determine the distribution of the link types it contains to a certain extent, including: statistics on the webpage size, text lengths and image sizes.

Two types of classifiers are used: SVM and decision tree. SVM can be used for both soft and hard classification. For the soft classification case, a distribution over all the class labels, rather than an absolute determination, is obtained for the link indicating the probability of level that the link belongs to each class, which is a more precise description of the link function. Decision trees are for hard classification only, but they naturally handle multi-class problems and perform much better in the imbalanced case, i.e. some classes cover only a minor part of the whole sample space.

3. EXPERIMENTS

We conduct some experiments to show the performance of our proposed link function classification algorithm. To insure the diversity of the testing data set, we performed Google searches with 10 popular terms and extracted all hyperlinks in top 100 results of each search. Finally 901 out of the 1000 results are available html pages and 57943 visible hyperlinks are extracted. All these links are manually labeled by their major function class. Table 1 summarizes the test data set.

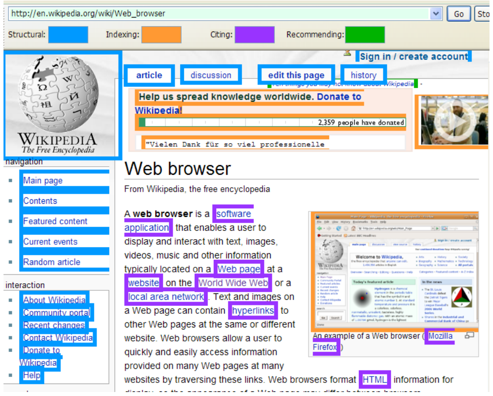
Table 1. Summary of the experiment data

Class	*Struct.*	*Indexing*	*Citing*	*Recom.*	*Advert.*
Num.	23509	10268	4673	17289	2204

Table 2. Accuracy of link function analysis algorithm

	SVM		D. T.		D. T. Non-spatial	
	P	*R*	*P*	*R*	*P*	*R*
Struct.	0.929	0.917	0.969	0.973	0.888	0.890
Index.	0.740	0.812	0.909	0.906	0.793	0.799
Citing	0.747	0.834	0.862	0.854	0.798	0.766
Recom.	0.905	0.842	0.954	0.952	0.876	0.878

The proposed SVM and decision tree classifier, as well as the decision tree based on only non-spatial features, are trained and tested. The *precision* (*P*) and *recall* (*R*) of 5-fold cross-validation for each classifier are reported in Table 2. The decision tree results are better than the SVM, which agrees with former discussions. But decision trees produce only hard labels. Most previous work dealing with link ranking use non-spatial features only, the experiment results show that the spatial features evidently improve the classification performance. The spatial features describe the actual visual representation of the link in browsers and are essential for human users to judge the link. A demo application was made upon the link function classification algorithm, and Figure 1 shows one of the classification results.


http://en.wikipedia.org/wiki/Web_browser

Structural: 🟦 Indexing: 🟧 Citing: 🟪 Recommending: 🟩

Help us spread knowledge worldwide. Donate to Wikipedia!

2,359 people have donated

"Vielen Dank für so viel professionelle

Web browser

From Wikipedia, the free encyclopedia

A **web browser** is a software application that enables a user to display and interact with text, images, videos, music and other information typically located on a Web page at a website on the World Wide Web or a local area network. Text and images on a Web page can contain hyperlinks to other Web pages at the same or different website. Web browsers allow a user to quickly and easily access information provided on many Web pages at many websites by traversing these links. Web browsers format HTML information for

An example of a Web browser (Mozilla Firefox)

Figure 1. Link function classification demo application

4. CONCLUSION AND FUTURE WORK

We have introduced the idea of link function classification and proposed an algorithm to analysis the link functions based on both vision and structure features, which is a simulation of how human users treat and judge the links. We plan to apply the link function classification on some of the popular link applications and to develop more learning and mining methods on the link sub-graphs produced as our future work.

5. REFERENCES

[1] Cai, D., He, X., Wen, J.-R. and Ma, W.-Y., Block-level link analysis. In Proc. of ACM SIGIR, 2004.

[2] Cai, D., Yu, S., Wen, J.-R. and Ma, W.-Y., VIPS: a vision-based page segmentation algorithm. Microsoft Technical Report, MSR-TR-2003-79, 2003.

[3] Lin, Z., King, I. and Lyu M. R.: PageSim: a novel link-based measure of web page similarity. In WWW 2006.

[4] Liu, N. and Yang, C. C.: A link classification based approach to website topic hierarchy generation. In WWW 2007.