Visual PageRank: Improving the Random Surfer Model Using Visual Features

Diligenti, Michelangelo and Kovačević, A., Miloš

Abstract–Standard techniques for estimating web page relevance neglect the information provided by the visual layout of the page. The most popular link-based technique, PageRank, uses a model of a random surfer to estimate the probability that he visits a page at any given time. This probability is assumed to be proportional to the relevance of the page. PageRank considers all outgoing links equally probable to be followed by the random surfer. Web designers use visual features to assign different degrees of relevance to the links and help users in browsing and selecting the information they need. Discarding this information makes the random surfer assumption more unrealistic and could have a strong negative impact in the accuracy of the predicted relevance estimations.

The proposed approach is based on the layout analysis that assigns to each link in a page a set of visual features. This set of features describes the visual appearance of the link and the logical context in which it appears. We created a model based on the real user observations to represent how humans select links based on their visual appearance. Finally, the user model and the visual information that we made available for each link are merged to estimate the probability that each single link will be followed in any given page. The experimental results showed that the visual PageRank improves the accuracy of the relevance estimation.

Index Terms—PageRank, user model, *visual analyses*

1. INTRODUCTION

COMPUTING the relevance of web pages is of fundamental importance for search engines which have to select a small set of results among a huge number of candidates that are matching a typical query. The average number of search terms provided by a user of a search engine is in the range of 1-2 terms (queries with more than five terms are rare). For such short queries millions of results are the norm, and assessing page relevance becomes fundamental in order to provide useful results to the users. Therefore a good scoring function is commonly considered to be the keystone for a successful search engine.

PageRank [8] is the most popular algorithm to estimate the relevance of' web pages. It exploits only the link-based information to achieve its goal. In particular, PageRank uses a model of a random surfer to estimate the probability that the surfer is at any given time visiting a page. This probability is assumed to be proportional to the relevance of the page. While PageRank considers all outgoing links equally probable to be followed by the random surfer, this assumption is very unrealistic. Web designers use visual features to assign different degrees of relevance to the links and to help users in browsing and selecting the information they need. Clearly, a big link in the upper part of a page has more relevance than a small link in the bottom and it should be weighted accordingly. .

Since pages are designed for humans, the meaning of the raw data obtained from a HTML source could be better understood if we could know the spatial relations among text, images and other objects. The goal of the presented research was to better model a real human user using the visual features of a page. In the proposed solution we employ a novel approach based on the layout analysis (obtained via an implicit rendering of the page), that assigns to each link a set of visual features. This set of features describes both the visual appearance of the link and the logical context in which it appears.

In the proposed algorithm, the rendering step computes the coordinates and relative positions of the objects in a page. Then, each page is represented as a hierarchical structure called *Visual Adjacency Multigraph,* in which nodes represent simple HTML objects like text and images. Directed edges connecting the nodes of the multigraph represent spatial relations between

Manuscript received May 4, 2009. (specify the date on which you submitted your paper for review.)

Diligenti. M. is with the Department of Informatics, University of Siena, Italy (e-mail: michi@dii.unisi.it).

Kovačević. M. (corresponding) is with the Department of Management, Technology of Building and Informatics, Faculty of Civil Engineering, University of Belgrade, Serbia (e-mail: milos@grf.bg.ac.rs)

the objects on the browser screen. The visual information contained in the multigraph allows defining heuristics for the recognition of common page entities such as vertical and horizontal link lists, titles and subtitles, and paragraphs of text.

Thus, visual analysis enables more accurate representation of the page contents, which splits the page into different logical portions. Moreover, at the end of the layout analysis step, each link is tagged with useful visual features like: size, color, position in the page, logical page portion to which it belongs (left or right menu, body, footer, header, etc.).

In order to be able to take advantage of the richer page (link) representation, it is needed to create a model of how real users select links based on their visual appearance. We decided to create such a model by recording the browsing behavior of real users. Finally, the user model and the visual information available for each link were merged to assign different weights to links, computed as the expected probability that the link will be followed by a user.

The outline of the paper is as follows: Section 2 describes the procedure of the visual layout analysis and defines heuristics for the recognition of logical groups in a web page. Section 3 presents a more general surfer model that is able to capture a more complex browsing behavior than the simple random surfer underlying PageRank. In Section 4, we present how visual information is embedded into the generalized PageRank model. Section 5 describes how the navigating behavior of users and their process of selecting links are modeled directly from users' observations. In Section 6, experimental results on a predefined datasets are shown. Finally, the conclusions are drawn in Section 7.

2. VISUAL LAYOUT ANALYSES

The main goals of the visual layout analysis are:

- Detecting the visual features of objects in a page. In particular, in this paper we will concentrate on links (with the corresponding anchor text), extracting the visual features as the color of the anchor text and its size. The size of a link is measured as the area on which the user can perform a click to follow it.
- Detecting the logical groups of objects in a page as they appear in a browser window, processing the corresponding textual file in an HTML format [7].

2.1 Logical groups of objects

Displayable objects are text, images, active-x controls, form controls and other. Each object is bounded with a boundary polygon of *n* vertices (where *n* is usually 4) and neighboring edges are always perpendicular. The logical groups of objects that can be taken into account strongly depend on the task at hand. For example, for the page relevance estimation problem that is addressed in this paper, a meaningful partitioning of a page could be to detect the portions of the page (and the corresponding links) that belong to:

- Body that is the central portion of the page where the most relevant information is presented. In some application like text classification, it could be possible to further subdivide the body into paragraphs.
- Vertical and horizontal menus (sequence of links). In particular, the vertical menus can be further subdivided into left and right menus, which have different semantic properties in a typical page.
- Header, the upper part of a page.
- Footer, the portion of a page at the bottom, which is often well visually distinct from the body and contains information which is non-essential to the rest of the page like: copyright information, information about the tool used to create the page or information about the author of the page, etc.

We emphasize that it is not possible to rely on HTML tags to determine whether a portion of a page belongs to a certain context. For example, some web designers could use the <.P> tag for labeling text and links that belong to a paragraph. But that assumption is highly inaccurate because other designers could use different tags to create the layout of a paragraph. An obvious example of misuse of tags is the usage of the <TABLE> one. Tables are commonly used (in 88% of the cases) to organize the layout of a page and the alignment of other objects, but not to organize tabular data [6]. It is not possible to rely on the proximity of text groups in a HTML source code because of previously mentioned <TABLE> tag and table nesting. Two text elements, which are close to each other in a source file, could be on the opposite sides of the screen. In order to simulate the visual human recognition the following problems have to be solved:

- Definition of an appropriate data structure that will reflect the positions of the objects and the spatial relations among them.
- Definition of a general recognition heuristics for each logical group of interest to be applied on the generated data structure.

The very first step in the visual layout analysis is to parse the page, and to construct the DOM tree that reflects the structure of the HTML source (nesting of container and inline tags/objects). A rendering procedure should be defined to calculate screen coordinates of the objects in a DOM tree such as pure text, links and images. The rendering procedure should imitate the behavior of popular

browsers. At the end of the first phase we obtain a coordinate tree in which the leaf nodes are represented as the coordinates of polygons embedding links, text boxes, and images. Each bounded polygon is represented as a set of four or more vertices defined with (x, y) screen coordinates (being the upper left corner of the screen the origin of the coordinate system). Details about the rendering process and the construction of the coordinate tree can be found in [7].

2.2 Visual Adjacency Multigraph

The recognition of logical groups of interest requires the understanding of spatial relations among the elementary constituents of a page. For example, when looking for vertical link lists it is needed to know if links are positioned in a vertical sequential row. Therefore it is needed to transform the coordinate tree from the rendering process into a more appropriate structure from which spatial relations among objects could be easily inferred. Before we define that structure more formally, we introduce four types of spatial relations among nodes of a coordinate tree. Let C be a coordinate tree and $p \in C$, $q \in C$ be two objects in the tree with bounded polygons C_p and C_q respectively. We introduce following definitions:

Definition 1: $p \leftarrow q$ (p is immediately left to q) if and only if (iff) the following statements are true:

1. $C_p \cap C_q = \varnothing$ (i.e. the polygons are not overlapping)
2. There exists at least one segment AB parallel to the x-axis connecting a point A $\in C_p$ and B $\in C_q$ (the intersection of the two projection of the polygons Over the v-axis is non zero).
3. There exists at least one segment AB for which rule 2 is valid such that for each $r \in C \setminus \{p, q\}$, AB $\cap C_r = \varnothing$ (the overlapping of the projections over the y-axis are not entirely covered by other objects).

Similarly,

Definition 2: $p \uparrow q$ (p is immediately before, or upper to q) iff the following statements are true:

1. $C_p \cap C_q = \varnothing$ (i.e. the polygons are not overlapping)
2. There exists at least one segment AB parallel to the y-axis connecting a point A $\in C_p$ and B $\in C_q$ (the intersection of the two projection of the polygons Over the x-axis is non zero).
3. There exists at least one segment AB for which rule 2 is valid such that for each $r \in C \setminus \{p, q\}$, AB $\cap C_r = \varnothing$ (the overlapping of the projections over the x-axis are not entirely covered by other objects).

The relations $\rightarrow$, (immediately right to) and $\downarrow$ (immediately after or down) can be defined symmetrically to the definitions above.

Definition 3: The *Visual Adjacency Multigraph* (VAM) of document d (represented as a set C of entries from a coordinate tree) is a directed multigraph with a set of nodes N and four sets of edges $E_{\leftarrow}$ $E_{\rightarrow}$ $E_{\uparrow}$ $E_{\downarrow}$. These sets have the following characteristics:

1. $N = |C|$ that is each object in the coordinate tree is associated to a node in the graph
2. $E_{\leftarrow} = \{(p,q) | p,q \in N: p \leftarrow q\}$. $E_{\leftarrow}$ contains all couples of objects p and q which are in $\leftarrow$ relation (all couples of objects which are horizontal immediate neighbors).
3. $E_{\rightarrow} = \{(p,q) | p,q \in N: p \rightarrow q\}$.. Symmetric definition to 2.
4. $E_{\uparrow} = \{(p,q) | p,q \in N: p \uparrow q\}$. $E_{\uparrow}$ contains all couples of objects p and q which are in p $\uparrow$ q relation (all couples of objects which are vertically immediate neighbors).
5. $E_{\downarrow} = \{(p,q) | p,q \in N: p \downarrow q\}$. Symmetric definition to 4.

For each couple (p, q) in $E_{\leftarrow}$, $E_{\rightarrow}$, $E_{\uparrow}$ or $E_{\downarrow}$, the VAM also stores the type of objects that are connected (links, text, images, etc.), their distance (in pixels) and a Boolean value describing if the polygons C_p and C_q containing the objects are vertically aligned for $E_{\uparrow}$ and $E_{\downarrow}$, or horizontally aligned for $E_{\leftarrow}$ and $E_{\rightarrow}$.

Figure 1 shows the whole process in which a web page is transformed into corresponding VAM.

2.3 Recognition of logical areas in a web page

Given a page p and its VAM(p), we want to label the links that belong to vertical link lists (and more specifically to left and right menus), horizontal link lists, the body of the page, the footer and the header of the page. Unfortunately, designers have different approaches to organize the same information to be displayed on the screen. Each designer can use different HTML tags or apply various cascading style sheets. Therefore we cannot directly rely on the information contained in the tags, except for rendering the page. After the rendering has completed, the positions, the spatial characteristics (like size, colors), and the relative spatial relations among HTML objects contained in the VAM are known.

In the proposed system a set of heuristics is used to discover the logical partitions from the VAM representing a page. As an example of heuristic used in the system, we describe the

heuristic applied to detect vertical menus.

Vertical Menu Heuristic: a vertical menu is detected using VAM as a path through the nodes in $E_{\uparrow}$ that are associated to links, that are aligned and that are closer than a given threshold.

It is clear that, given a VAM, it is easy to create other heuristics for detecting different partitions that could be of interest for other applications. Refer to [7] for a more extended description of these heuristics.

A big advantage of the approach is its flexibility and generality: having information about relative relations among objects, no matter where they are positioned, it is possible to easily create new heuristics for extending the recognition to other logical groups. This allows the application of the same framework to other problems like: page classification, focus crawling, recommending systems, page segmentation, etc. Probably the weakest point of the approach is the complexity of the multigraph construction process. In the current implementation the entries of the coordinate tree are processed, and for each new entry we check all previously encountered ones for spatial relations. Therefore, the complexity is $O(n^2)$, where n is the number of entries in the related coordinate tree (in a typical page less than 100 entries are included in the tree). There is a lot of space to optimize the implementation of rendering and recognition tasks, but that was left for future work.

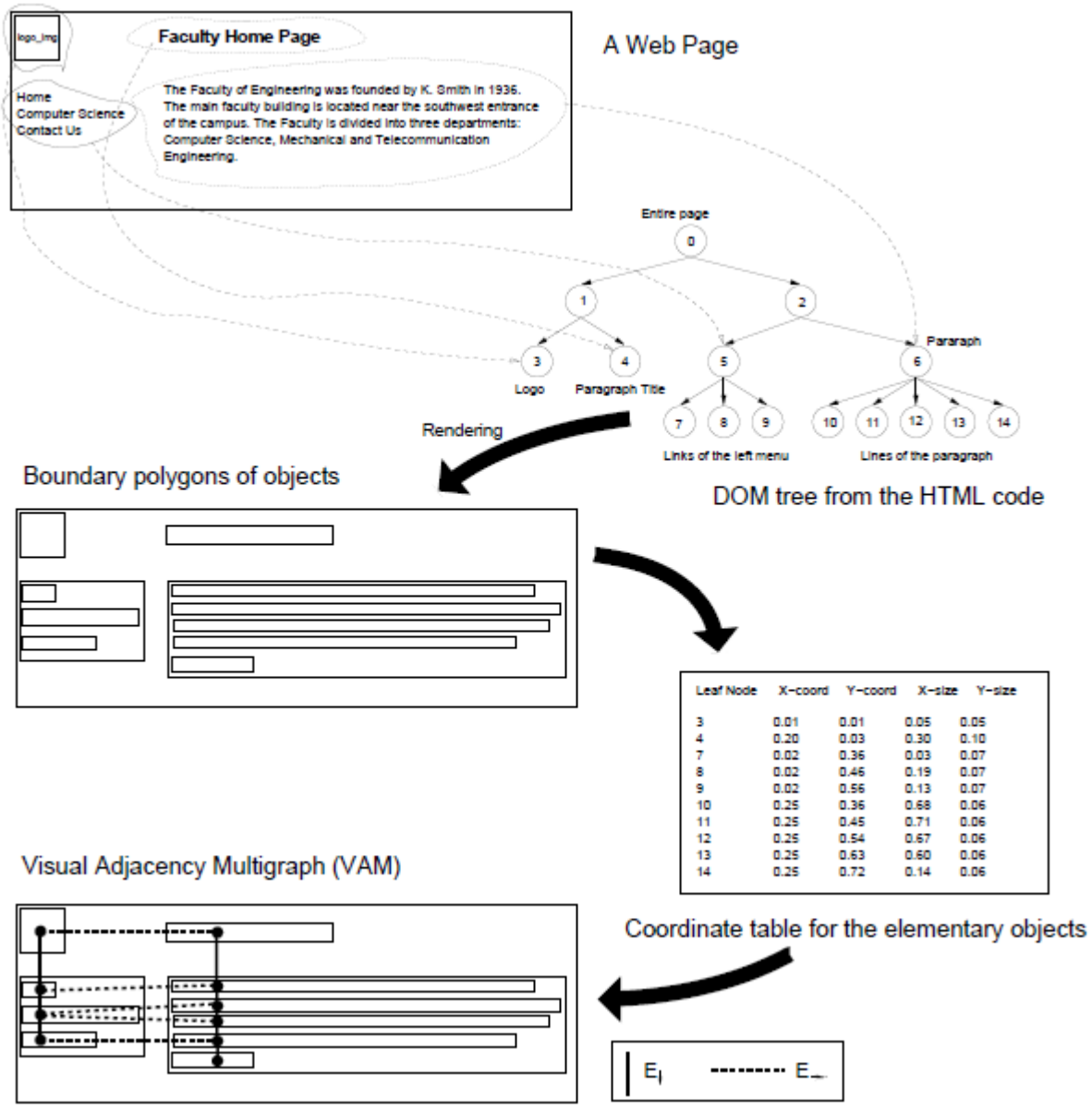

***Figure 1*: Given a web page, we render the page starting from the DOM tree extracted from the HTML source code. The output of the rendering module is the set of coordinates of the boundaries of the polygons containing the objects in the page. The Visual Adjacency Multigraph (VAM) is then extracted from the boundary polygons. Each edge of the VAM also includes information about the distance between the connected objects and if they are aligned. For sake of readability, this**

information is not displayed in the figure.

In [1] it is observed that a common surfer expects to see certain kinds of objects such as menus and banners in predefined screen areas of a page. This could allow sorting the elementary objects and restricting the areas of the screen in which to search specific logical objects resulting in a dramatic speed up of the process.

3. EXTENDED RANDOM SURFER MODEL

PageRank introduces a notion of page authority which is completely independent on the page content, whereas the authority only emerges from the topological structure of the Web. In PageRank, the authority is similar to the notion of citation in the scientific literature. In particular, the authority of a page *p* depends on the number of incoming hyperlinks (number of citations) and on the authority of the page *q* which cites *p* with a forward link. Moreover, selective citations from *q* to *p* are assumed to provide more contribution to the score of *p* than uniform citations. Hence, the PageRank x_p of *p* is computed by taking into account the set of pages *pa*[*p*] pointing to *p*. According to [8], the PageRank is computed using the following set of recursive linear equations:

$$x_p = \lim_{t\to\infty} x_p(t) = d \sum_{q\in pa[p]} \frac{x_q(t-1)}{h_q} + \frac{1-d}{|W|} \qquad (1)$$

$$\forall p \in W$$

Here, *W* is the set of pages in the considered Web, and $d \in (0,1)$ is a *damping factor* and h_q is the *hubness* of *q*, that is the number of hyperlinks contained in *q*.

The PageRank models a random walk on the Web graph of a surfer allowed to perform only two actions at each time step: the surfer jumps to a new random page with probability 1 - *d* or she/he follows one link from the current page with probability *d*. All these values are considered to be independent on the page *p*. Given that a jump is taken, its target is selected using a uniform probability distribution over all the I*W*I Web pages. Finally, the probability x(*p* I *q*, *I*) of following the hyperlink from page *q* to page *p* does not depend on the page *p* i.e. x(*p* I *q*, *I*) = α_q In order to meet the normalization constraint, α_q=1/h_q where the hubness of page *q*, h_q = *ch*(*q*), is the number of links exiting from page *q* (the number of children of the node q in *W*). This requirement cannot be met by sink pages, i.e. the pages which do not contain any link to other pages. In order to keep the probabilistic interpretation of PageRank, all sink nodes must be removed and their scores computed only after the model has converged. Typically, the initial page scores $x_p(0)$ are uniformly set to 1/I*W*I. Using well studied results of Markov chains theory [9], it can be easily shown that equation (1) converges. At convergence, the PageRank x_p of page *p* represents the probability that the random surfer will be located at *p* at any time step.

4. VISUAL PAGERANK

The main limitation of PageRank arises from the assumptions made about the underlying random surfer. Such assumptions are unrealistic and do not accurately model how a real surfer browses the Internet. In particular, it is strongly unrealistic that a user uniformly jumps to a random page when it decides that the current page does not lead to any further useful information. It is more likely that he will restart his surfing from some authoritative page or search engine. In order to improve the random surfer scheme, it is possible to use a biased distribution of the pages where the surfer will land after a jump [3].

Moreover, the surfer model assumes that all out-links of a page are equally likely to be followed. This is clearly unrealistic since users are very good in selecting specific links that they think to be useful. The surfer model could be easily extended to take into account a bias on the out-link probability [3, 4]:

$$x_p(t) = d \sum_{q\in pa[p]} x(p \mid q, I) x_q(t-1) + \frac{1-d}{|W|} \qquad (2)$$

$$\forall p \in W$$

However, whereas the model can be easily extended, it is not clear which semantic should be assigned to the link weights x(*p* I *q*, *I*).

In [3], this extended PageRank surfer model is used to create a topic-specific (or vertical) page rank, where the link strength *x*(*p* I *q*, *I*) is assigned to be proportional to the score assigned to the target page *p* by a classifier-by-topic. However, it is not clear how to bias the weights for non-vertical search.

Following the assumption that modeling a more intelligent and "real" surfer leads to a more accurate ranking of the pages [4], we propose to use the visual information to emulate the complex analysis that a user pursuits when looking at a document displayed on the screen. Since links have different visual characteristics, a user is influenced during the selection process by the link size, its color, position in the page, by whether the link is an image, and by many more factors. Using the visual approach, we aim at estimating the probability that a real user, looking at the page,

would click on each single link on it.

5. ESTIMATING THE PARAMETERS OF USER MODEL

Since we aim at modeling the user behavior, we first tried to understand how real users are influenced by visual features and, then, to estimate model parameters directly from these observations. Let the links be subdivided into the following categories:

- Link with images: only reasonably big images with size greater than 1% of the screen were considered in this context.
- Links with emphasized anchor text. This category includes links with bold, capitalized or underlined anchor text or anchor text with font size at least one point bigger than the average text in the page
- Links with non-emphasized anchor text. These links do not "visually" emerge from the rest of the page.
- Other links.

Moreover, links are marked as belonging to the header, footer, body, left: and right menu of the page as described in section 2. Clearly, the above classification discards many other possible link categories. This simplification was needed to keep the number of the parameters of the user model to a reasonable level. We observed a set of 21 Internet users during their common search and browsing activities on the Web and we recorded their clicks. In particular, each user was recorded during 5 different information searching sessions for a total of 63 sessions and more than 2500 clicks.

	Header	Footer	Body	Right Menu	Left Menu
% clicks on links from	19.5	10.5	28.4	10.8	30.8

(a)

	Images	Emphasized Anchor Text	Standard Anchor Text
% clicks on links of type	23.8	13.0	6.1

(b)

***Table 1*: (a) percentage of clicks performed by the user on the links in each considered partition of a page. (b) for each link type, we recorded the percentage of elements that received at least one click for one user.**

	Header	Footer	Body	Right Menu	Left Menu
Link Strength	+0.06	-0.124	+0.101	-0.105	+0.122

(a)

	Images	Emphasized Anchor Text	Standard Anchor Text
Link Strength	+0.124	+0.09	-0.0248

(b)

***Table 2*: (a) visual strength of a link in each page partition. The strength is measured as the deviation of the observed user click probability from the baseline (the case of' a random surfer who randomly selects links without looking at the page). (b) The computed visual strength of each link category.**

For each visited page, we recorded the numbers of clicks on links from each considered category. In Table 1, there are reported the percentage of user clicks on a specific portion of the page and the percentage of clicked elements for each link category, respectively. The tables show that users tend on average to click more on image links and on links with emphasized anchor text. As expected, users click more likely on links in the body or left menus of a page. On the other hand, the user is less likely to click on links that are associated to standard anchor text (which is not visually emerging) and on links located in the footer and right menu of the page. These probabilities can not be directly used as weights of the user model since user behavior depends on the actual page, i.e. the user will not click on a link in the left menu if such menu is not present in the page. Similarly, a user could be very likely to click on links with emphasized anchor text simply

because such links are more commonly found in a given Web document. As a conclusion, simply counting how many times the users click on a link of a specific category is not a good guess to measure how much a link category attracts the users. To solve this problem, the weights for a link of a specific category must be adapted to the average number of links of that category that are present on the Web pages. In order to do that, we define the visual strength S_t of a link type t as:

$$S_t = \frac{c_t}{C} - \frac{n_t}{N}$$

where c_t is the number of clicks on links of category t recorded during all sessions, C is the overall number of clicks recorded during all sessions, n_t is the number of links belonging to category t that appeared in the pages accessed during all sessions, and N the total number of links in all displayed pages. S_t represents the visual strength of each link type, measured as the deviation of the observed click probability (posterior probability) from the expected a-priori probability of that link category.

Obviously, a value S_t grater (or smaller) than zero means that specific link category is more (or less) likely to be clicked than expected when using a uniform distribution over the links. In general, the higher the value of S_t, the more likely the link will be selected by the user with a click and, therefore, the more bias should be assigned to the link during the random walk. Table 2 shows visual strengths for each considered link category.

We considered the bias due to the link type to be independent from the bias due to the position of the link in the page. This "naive" assumption is clearly not true, but it was needed to limit the number of parameters that had to be estimated while recording and analyzing the users' behavior.

For each page q each outgoing link to a target page p is classified according to its type (image, emphasized anchor text, standard anchor text) and its association to a given logical group (body, left/right menus, footer, header). Then the weight $x(p \mid q, l)$ of the link from page q to page p (the probability of the surfer to follow a link to page p given that he is located in q and he will not jump) is computed as:

$$x(p \mid q,l) = \frac{1}{h_q} + S^1_{q\to p} + S^2_{q\to p}$$

where $S^1_{q\to p}$ and $S^2_{q\to p}$ are the strengths due to link category and link position for the considered link, respectively. When $S^1_{q\to p}$ and $S^2_{q\to p}$ are equal to zero (i.e. there is no evidence that the visual features affect the user behavior), equation (2) becomes the standard PageRank equation. These parameters are then normalized for each page q to sum up to one and are then inserted into equation (2) to compute the Visual PageRank.

Other interesting information emerged from the observations of user browsing behavior. For example, we noticed that users are less likely to click on advertisement images (like banners) than on other image links. On the other hand, users tend to click more on company logos. Even using visual information, our system was not able to perform the sophisticated page analysis needed to distinguish between advertisement image links and company logos. Thus, we decided to discard this potentially useful information.

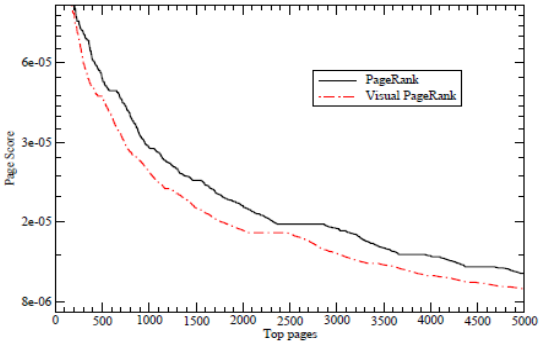

***Figure 2*: The score distribution of the set of top 5000 pages: the visual rank yields a less smooth and more "discriminative" score distribution.**

6. EXPERIMENTAL RESULTS

We performed a set of experiments to compare how the Visual PageRank differs with respect to the standard PageRank. Using the focus crawler described in [2] we downloaded 150.000 documents related to the topic "wine". We then computed the PageRank and the Visual PageRank using the surfer parameters estimated as described in the previous section, on this focused portion of the Web.

We expected the visual information to prevent the rank from flowing through not important links, while increasing the rank flow through the most relevant links in each single page. A first confirmation of this effect is shown in the plot in figure 2, where it is plotted the relevance of the top 5000 documents for the Visual and standard PageRank. The Visual PageRank is less smooth assigning higher values to a smaller subset of pages.

Table 3 shows links of 20 pages with the highest score. Using the Visua] PageRank all top results were relevant for the topic "wine", whereas using standard PageRank some pages that are not authoritative for the considered topic showed up in the first positions, i.e. "www.microsoft.com/ie/logo.asp" and

"www.microsoft.com/windows/mediaplayer/download". Even if Visual PageRank is not focused on any specific topic (the content of the page is never taken into account), it is able to not assign a high score to these pages, even if they are linked from a high number of pages. This happens because such links are usually not part of the page core and they are not intended to be immediately seen by the user.

7. CONCLUSION

Visual information contained in a web page is usually discarded. However, this information is very important for a better understanding of the content of web documents. In this paper, the visual information was extracted for any given page and used to predict the behavior of a real user when visiting that page. This prediction has been shown to be useful to generate an improved random surfer model and, as a consequence, a more precise ranking function for search engines.

In spite of the improvement, there is still a big gap between the complex interaction that a real user performs with a page and the simple user model that was employed in our experiments. For example, our experiments showed that real users are less likely to click on an advertisement image links (like banners) than on other image links. On the other hand, users tend to click more on company logos. Even using the visual information, the proposed system is, for the moment, unable to perform the sophisticated page analysis that could allow distinguishing between advertisement image links and company logos. Another possible improvement is to make the rendering module to take into account advanced features like layers and style sheets which are actually ignored, discarding some useful visual information for the most complex page layouts.

We also plan to apply the proposed framework to focused versions of PageRank like that proposed in [5, 8].

Finally, we plan to test the proposed ranking algorithm on a bigger dataset and to measure the ranking accuracy from the feedback of real users.

Please note that the visual approach presented in this paper is very general and it could be used to improve many other applications like page classification and clustering, information extraction, focus crawling and others.

ACKNOWLEDGMENT

We thank Chiara Pintucci from the "Dipartimento di Scienze della Comunicazione" of the University of Siena who performed the user observations that allowed to create the user model used in the paper.

PageRank	Visual PageRank
www.winerelease.com/useragreement.htm	www.winespecialist.com
www.winespecialist.com	www.wineathomeit.com
www.wineathomeit.com	www.connollyswine.co.uk
www.singlesites.com	wineportal.itswine.com
www.closmimi.com	www.connollyswine.co.uk/crt_Cart.asp
www.erobertparker.com/support/...	www.sbwines.com
www.thevirginiacompany.com	forum.wine.co.za/display...
www.connollyswine.co.uk/crt_Cart.asp	www.thevirginiacompany.com
www.connollyswine.co.uk/prodsearch.asp	www.hawleywine.com
www.erobertparker.com/subscriptions/	www.wineanorak.com/taste.htm
www.wineanorak.com/taste.htm	www.ravenswoodlane.com.au
microsoft.com/windows/mediaplayer/...	www.montrosechina.com
www.liquidasset.com	annex.winehouse.com
macromedia.com/shockwave/download/...	www.liquidasset.com
www.adremote.timeinc.net/click.ng...	www.wineanorak.com
www.montrosechina.com/include_html/...	www.clicquot.com
annex.winehouse.com/past.reviews.html	www.cellarexchange.com
www.microsoft.com/ie/logo.asp	www.santabarbara.winecountry.com

***Table 3*: Top 20 score pages for the dataset "wine" when using PageRank and Visual PageRank. Note two irrelevant links from WWW.MICROSOFT.COM website when classical PageRank is used.**

REFERENCES

[1] L. M. Bernard. Criteria for optimal web design (designing for usability). Technical report, Wichita State University, available online at http://psychology.wichita.edu/optimalweb/position.htm, 2001.

[2] M Diligenti, F. Coetzee, S. Lawrence, L. Giles, and M. Gori. Focus crawling by context graphs. In *Proceedings of the International Conference on Very Large Databases*, 11-15 September 2000, Cairo, Egypt, pages 527-534, 2000.

[3] M. Diligenti, M. Gori, and M. Maggini, A unified probabilistic framework for web page scoring systems. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16(1):4-16. January 2004.

[4] P. Domingos and M. Richardson. The intelligent surfer: Probabilistic combination of link and content information in pagerank. *Advances in Neural Information Processing Systems,* 14:1441-1448, 2002.

[5] T. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th World Wide Web Conference (WWW11)*, Honolulu, Hawaii, May 2002.

[6] F. James. Representing structured information in audio interfaces: A framework for selecting audio marking techniques to represent document structures. Technical report, Ph.D. thesis. Stanford University, available online at http://www.pcd.stanford.edu/frankie/thesis/, 200l.

[7] M. Kovacevic, M. Diligenti, M. Gori, and V. Milutinovic. Recognition of common areas in web page using visual information: a possible application in a page classification. In *Proceedings of International Conference on Data Mining (ICDM)*, pages 250-258. Maebashi City, Japan, 2002. IEEE Press.

[8] L Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bringing order to the web. Technical report, Computer Science Department, Stanford University, 1998.

[9] E. Seneta. *Non-negative matrices and Markov chains*, Springer Verlag, 1981.