Web Mining and Online Visibility

Nadine Schmidt-Mänz and Wolfgang Gaul

Institut für Entscheidungstheorie und Unternehmensforschung,
Universität Karlsruhe (TH), 76128 Karlsruhe, Germany

Abstract. In order to attract web visitors via the internet online activities have to be "visible" in the net. Thus, visibility measurement of web sites and strategies how to optimize Online Visibility are important. Here, web mining helps to define benchmarks with respect to competition and allows to calculate visibility indices as predictors for site traffic.
We use information like keyword density, incoming links, and ranking positions in search engines to measure Online Visibility. We also mention physical and psychological drivers of Online Visibility and describe the appropriateness of different concepts for measurement issues.

1 Introduction – "Why measurement of online visibility?"

Search engines appear to be very important to reach new visitors of web sites, because nearly 80% of all internet users find new web sites with the aid of search engines (Fischerländer (2003)). Results by Johnson (2002) also show that more active online shoppers tend to search across more sites and that the amount of online search is actually quite limited when internet surfers already have a special portfolio of web sites.
Therefore, it is very important to observe what could be called Online Visibility of web sites (see Drèze and Zufryden (2003) who have defined Online Visibility as the extent of presence of a brand or a product in the consumer's environment, e.g. by means of links from other web sites, online directories and search engines).
We suggest a measure called GOVis (Gage of Online Visibility) to keep track of the Online Visibility of web sites and to measure the success (unsuccessfulness) of conducted web site optimization.
In section 2 we shortly describe the web as a graph and focus on facts about human online searching and surfing behavior. We explain our measure of Online Visibility and main drivers to influence this phenomenon in section 3 while conclusions and some managerial implications are given in section 4.

2 (Human) Online search in a changing webgraph

The structure of the web is often compared with a haystack in which one tries to search for and find the needle. If the web is modeled as a directed

graph, the addition of new vertices and edges and the omission of old ones cause changings of the graphical structure.
Researchers try to build subgraphs of the complete underlying web graph and develop models to interpret evolving views on this dynamically altering net.
But what do persons really see when they are searching the web? They only get sub-subwebgraphs corresponding to their search efforts and requests which describe just static parts of the underlying situation.
In order to understand (human) online searching and surfing behavior and to derive managerial implications for web site owners adequate measures with respect to the underlying phenomena are needed.

2.1 The web as a graph

A web site consists of pages connected in a certain way. Their link structure can be described by the associated site graph. The web consists of sites and hyperlinks between certain pages within the same site (site subwebgraph) and across different sites. The pages can be seen as vertices in a directed graph and the hyperlinks as directed edges. If one tracks the web, one gets from vertice to vertice by following the directed edges. In the end one has information based on the structure of the tracked subgraph represented by its adjacency matrix. Given this information it is possible to calculate measures to characterize this subwebgraph.
Here, some facts about this microscopic view on the web have to be mentioned (Barabasi and Albert (1999), Broder et al. (2000)):
The average distance (also referred to as diameter) as number of links to get from any page to any other is about 19, if a path exists. The distribution of

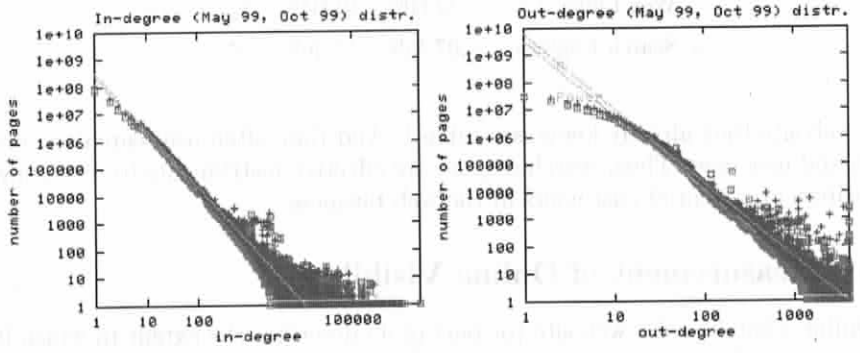

Fig. 1. Distribution of in- and outgoing links of pages

incoming and outgoing links of pages in the web is shown in figure 1 (Broder et al. (2000)) with the number of incoming and outgoing links on the x-axis

and the number of pages with corresponding in- and out-links on the y-axis, both depicted on logarithmic scales.We have used this shape to model the function $f(Z_L)$ in section 3.3.

2.2 (Human) Online searching and surfing behavior

Three main studies can be mentioned that report on online searching via web search engines by analyzing query logs: The Fireball, the Excite, and the AltaVista study (Hölscher (1998), Jansen et al. (2000), Silverstein et al. (1999)).
Conclusions of all three studies are nearly the same. The AltaVista study is based on the largest data set: one billion queries submitted to the main search engine over a 42-days period.
Facts about human online searching behavior corresponding to the AltaVista study can be summarized as follows: Nearly 77.6% of all query sessions consisted of only one request. 85.2% of the searchers examined only one result screen per query (7.5% two and 3.0% three screens). The average number of terms in a query adds up to 2.35 ($\sigma = 1.74$) and that of operators in a query to 0.41 ($\sigma = 1.11$). According to the total number of queries, 63.7% occurred only once. The most popular query was "sex" with an appearance of 1,551,477 times. This equals 2.7% of the total number of non-empty queries in the study.
People become also more efficient in using the web by navigating directly to

Table 1. Global Internet Usage (WebSideStory (2003))

Referral Type	**2002**	**2003**	trend
Direct Navigation	50.21%	65.48%	↗
Web Links	42.60%	21.04%	↘
Search Engines	07.18%	13.46%	↗

a web site they already know, see table 1. And they often use search engines to find new ones. Thus, search engines are effective instruments to reach new visitors or potential customers in the web business.

3 Measurement of Online Visibility

Online Visibility of a web site (or part of it) describes the extent to which it is recognizable or findable via normal searching strategies of web users.
Based on knowledge, e.g., about the link structure of the web as a graph, the functioning of ranking algorithms of search engines such as PageRank (Brin and Page (1998), Kleinberg (1999)), and human searching and surfing behavior, several impacts on Online Visibility can be defined.

3.1 Main drivers of Online Visibility

Online Visibility has to be composed of different visibility parts as, e.g., visibility via links from other web sites, visibility via listings in online directories, and visibility via search engines, to mention just the most important ones. Some information of this kind is already used by search engines within their strategies to place the most important web pages on top of corresponding listings. All in all two main kinds of drivers of Online Visibility can be identified:

1. Psychological Drivers of Online Visibility: This means that human online searching and surfing behavior and ways how humans interact with the internet or with search engines (e.g., only the first three result pages of search engines are normally inspected by browsing individuals) have to be taken into consideration.
2. Physical Drivers of Online Visibility: Physical drivers are such as links to a web site, banner ads, listings in search engines or directories etc.

Both psychological and physical drivers cause differences with respect to Online Visibility. To determine the real impact on Online Visibility one would have to subtract all overlappings from different visibility parts. For this reason, it is difficult to determine a precise measure. However, one can approximate a measure that takes the main phenomena mentioned into account.

3.2 Web data used for our sample

We used the Google search engine for data collection, because Google holds 73.4% of the share of the search engine market (percent of search requests), followed by Yahoo! with 5.5% (Webhits (2004)).

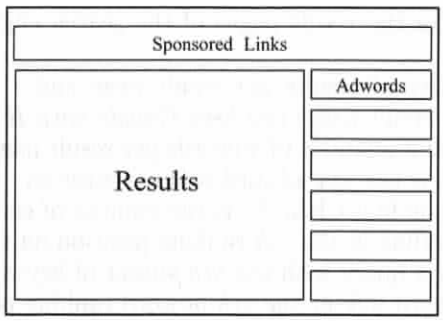

Fig. 2. Conventional result window of Google

If one considers a conventional result window of the Google search engine as it is shown in figure 2, one sees on the right hand side the so-called "Adword

Area", in the middle the result list of Urls corresponding to the search request, and up to two so-called "Sponsored Links" on top of the result screen. For our measurement of Online Visibility we keep track of the appearance of the Urls of interest in the result list and in the adword area for appropriate requests characterized by their keywords.
Additionally, we considered the AltaVista search engine for determining the number of incoming links as in AltaVista it is possible to exclude links from the home domain (link:www.xyz.com -host:www.xyz.com).
We excluded the measurement of Online Visibility in directories, on portals, in chat rooms or banner ads, etc. The reason is that it is not possible to measure OV, e.g., in directories in an impartial way (alphabetical order) and to take changes of banner ads into account without a huge amount of data from other webmasters (see, e.g. Drèze and Zufryden (2003) who incorporated expensive and time consuming information retrieval methods to calculate their measure, which is static and only a snap shot based on a selective situation in the web).

3.3 The measure GOVis

Obviously, there are many ways to try to formulate an Online Visibility measure but based on the reasons mentioned before our approach

$$GOVis(L) = \sum_{k=1}^{\sum_{n=1}^{N}\binom{N}{n}\cdot n!} \left[\alpha \cdot \sum_{p=1}^{2}\sum_{r=1}^{R}\frac{1}{e^{p-1}} \cdot X_{kpr} + \beta \cdot \sum_{p=1}^{2}\sum_{a=1}^{A} Y_{kpa}\right] + \gamma \cdot f(Z_L)$$

is a fast and cheap method and independent of third party data. Additionally, consecutive investigations can be performed. Here

* $\mathcal{K}$ is a set of interesting keywords for a query, with $\mid \mathcal{K} \mid = N$ (normally $N \leq 3$), $\sum_{n=1}^{N}\binom{N}{n} \cdot n!$ is the quantity of all ordered subsets of $\wp(\mathcal{K}) \setminus \{\emptyset\}$ and k is the kth subset of keywords with which a query in *Google* could be performed,
* p is the depth of the result pages of the search engine used (normally depth $p \leq 2$),
* R is the quantity of results per result page and r is the rth ranking position on the result pages (we used *Google* with $R = 10$),
* A is the maximum quantity of adwords per result page (*Google* standard is $A = 8$) and a is the ath adword ranking position,
* L is the corresponding URL, Z_L is the number of corresponding fan-ins,
* h_{kpr} is the hyperlink at the rth ranking position on the page with depth p by generating a query with the kth subset of keywords,
* w_{kpa} is the adword link at the ath adword ranking position on the page with depth p by generating a query with the kth subset of keywords,
* $X_{kpr} = \begin{cases} 1, & h_{kpr} \text{ links to } L \\ 0, & \text{otherwise} \end{cases}$ $\quad Y_{k1a} = \begin{cases} 1, & w_{kpa} \text{ links to } L \\ 0, & \text{otherwise} \end{cases}$
* $\alpha + \beta + \gamma = 1$ (these parameters help to adjust overlappings),
* and $f(Z_L)$ is a step function based on figure 1.

3.4 Results

We examined different branches: e.g., online book stores, erotic service web sites, automobile, and nonprofit web sites. We also observed "trendy" web sites such as the German home page of the movie "Lord of the Rings". The number of incoming links changed dependent on the branches observed (the number of incoming links of one book store decreased by 1,500 whereas the number of incoming links of erotic and nonprofit web sites were static in the last month (March 2004)). Incoming links of "trendy" web sites alter in accordance with the interest given to these web sites by press or television (e.g., the number of incoming links of the "Lord of the Rings" web site reincreased again after the academy awards of 11 "Oscars" in March 2004 up to 4,099 and fell down to 216 in April 2004).
The function $f(Z_L)$ that we used is based on the findings presented in figure 1, relative to $\log Z_L$, and absorbs changes in the number of incoming links to some extend (e.g., if a web site has already "many" incoming links, it doesn't matter if it gains or looses "some").
Figure 3 shows GOVis results of book stores ($\mathcal{K} = \{$dvds, roman, bestseller$\}$)

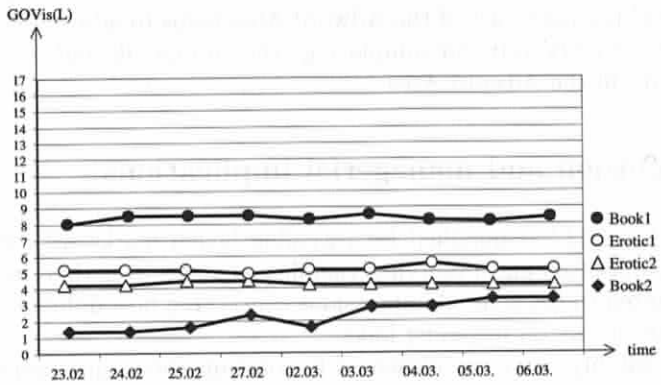

Fig. 3. $GOVis(L)$ for different web sites

and erotic service web sites ($\mathcal{K} = \{$erotik, sex, porno$\}$); $\alpha = 0.39$, $\beta = 0.01$, and $\gamma = 0.6$ was selected according to a scenario based on the numbers of table 1 (Take $\alpha \approx 13.46/35$, $\beta \approx 0.5/35$, and $\gamma \approx 21.04/35$).
One sees that the outcomes of $GOVis(L)$ for erotic web sites are very close to each other. Based on our measure managers of these sites can evaluate how their activities influence Online Visibility compared to salient competitors. For book stores we could select two example which demonstrate how different Online Visibility can be. Book1 is on an upper level of Online Visibility, because this web site has good listings in Google for nearly every keyword combination based on the chosen set $\mathcal{K}$ and has also many incoming links. Book2 can now try various activities and observe how it "best" (on-/offline

marketing campaigns) can approach this competitor (and, indeed, between February 23 and March 6, 2004, the GOVis measure has increased by two units). With the help of GOVis one can measure the "own" Online Visibility, but can also make online competitors visible to search for best practice examples of web sites and to derive hints for successful actions. Table 2 shows the

Table 2. Appearance of URLs for different branches from April to June 2004

Branch	#Urls in Total	#Urls in Results	#Urls in Adwords	% of Capacity of Adword Area
Book	502	377	117	48,9%
Erotic	1920	1058	853	91,55%
Automobile	565	337	215	99,6%
Nonprofit	1032	908	118	16%

appearance of Urls for different branches in online business. The appearance of URLs and the used part of the Adword Area helps to adjust the choice of α, β, and γ for GOVis. In our sample, e.g., the automobile and erotic branch is very active in the Adword Area.

4 Conclusion and managerial implications

In total, $GOVis(L)$ is qualified for revealing bench marks with respect to possible competitors and observing visibility changes over time in the web. It is also suited to get general impressions concerning how different branches use adwords or rely on incoming links.
However, visibility has to be measured in constant short time periods to get a deeper understanding of the rate with which the WWW or, more precisely, subwebgraphs are changing and how certain web activities influence the GOVis measure.
Based on the money spent on the optimization of web sites with respect to Online Visibility, one can observe the success (unsuccessfulness) of special arrangements with the help of GOVis. Although GOVis is only one suggestion to measure Online Visibility with the help of content visibility, adwords visibility, search engine visibility, and visibility based on incoming links (which, however, appear to be the most important instruments to account for Online Visibility), some managerial implications are obvious:
To improve Online Visibility it is important to follow different strategies. One question is, how the link structure of the web site is built up, and another one, how many links are pointing to different pages. An obvious strategy to generally improve the ranking in search engines is to be listed in online

directories. And, because there is an impact of the order of keywords in a query, the way of ordering the content of web pages has to be considered for long-time optimization of a corresponding web site. At first, however, web site owners have to find out which keywords are relevant for online searchers and web site content. For example, it is possible to observe the log files of corresponding web sites to detect search engine referrals including important keywords of searching persons. Another possibility is to sift through online keyword databases to compare already used keywords with descriptions or text content of the web site of interest to meet customer needs with respect to content, special topics or product descriptions. And if one detects potential competitors with GOVis, it is possible to analyze the sites of these competitors to find out whether their web appearance works better than the own one.

References

BARABASI, A. and ALBERT, R. (1999): Emergence of Scaling in Random Networks. *Science, 286, 509–512.*

BRIN, S. and PAGE, L. (1998): The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the 7th International World Wide Web Conference, WWW7, 107–117.*

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A. and WIENER, J. (2000): Graph Structure in the Web: Experiments and Models. *Computer Networks, 33, 309–320.*

DRÈZE, X. and ZUFRYDEN, F. (2003): The Measurement of Online Visibility and its Impact on Internet Traffic. *Journal of Interactive Marketing, 18(1), 20–37.*

FISCHERLÄNDER, S. (2003): Websites Google-gerecht - Ganz nach oben.*iX 08/2003, 84–87.*

HÖLSCHER, C. (1998): How Internet Experts Search for Information on the Web. In: H. Maurer and R.G. Olson (Eds.): *Proceedings of WebNet98 - World Conference of the WWW, Internet and Intranet.* AACE, Charlottesville, VA.

JANSEN, B., SPINK, A. and SARACEVIC, T. (2000): Real Life, Real Users, and Real Needs: A Study Analysis of User Queries on the Web. *Information Processing and Management, 36, 207–227.*

JOHNSON, E.J. (2002): On the Depth and Dynamics of Online Search Behavior. *Department of Marketing, Columbia Business School, Columbia University.*

KLEINBERG, J. (1999): Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM, 46(5), 604–632.*

SILVERSTEIN, C., HENZINGER, M., MARAIS, H. and MORICZ, M. (1999): Analysis of a Very Large AltaVista Query Log. *SIGIR Forum, 33(1), 599–621.*

Webhits (2004): Web-Barometer, Nutzung von Suchmaschinen, www.webhits.de, last visited April 14, 2004.

WebSideStory (2003): Search Engine Referrals Nearly Double Worldwide, According to WebSideStory, March 2003, www.websidestory.com, last visited March 18, 2004.