Adv Data Anal Classif (2011) 5:113–128
DOI 10.1007/s11634-011-0088-5

REGULAR ARTICLE

# Web page importance ranking

**Wolfgang Gaul**

Received: 18 October 2010 / Revised: 24 February 2011 / Accepted: 1 March 2011 /
Published online: 1 April 2011


**Abstract** An approach is proposed that uses a set of interesting Web pages as starting point for a minimum walk algorithm to provide recommendations of additionally important Web information within a m-clicks-ahead situation. A discussion of known page importance ranking techniques as well as examples of the application of the new algorithm show that Web link structure dependent approaches should be enriched by considerations as to how the analysis of additional data and the use of suited support tools can be incorporated. These considerations include aspects as, e.g., personalization, query dependence and topic sensitivity of the underlying pages, the dynamic nature of the Web, as well as the possibility to perform calculations online.

**Keywords** Page ranking · Web link structure analysis · Content similarity of Web pages · m-Clicks-ahead recommendations

**Mathematics Subject Classification (2000)** 68M11 · 68P10 · 68U15 · 68U35 · 68W27 · 90B60

## 1 Introduction

Finding, structuring, and analyzing Web information and understanding the behavior of internet users are important prerequisites for an area, called Web intelligence enhancement, that is concerned with establishing advantageous strategies derived from Web data analysis results and creating products and services based on newest electronic storage, evaluation, and interchange possibilities.

W. Gaul (✉)
Institut fuer Entscheidungstheorie und Unternehmensforschung,
KIT, Postfach 69 80, 76049 Karlsruhe, Germany
e-mail: wolfgang.gaul@kit.edu

the starting node and $j$ the end node of the directed edge $e = (i, j)$. Assume that for each pair of nodes at most one directed edge exists. Then, $A = (a_{ij})$ with $a_{ij} = 1$, if $(i, j) \in E$, and $a_{ij} = 0$, otherwise, is called adjacency matrix and describes the link structure of the graph. Often, one has a valuation $v : E \to I\!R$ with values $v_{ij}$ that allow for, e.g., capacities, costs, lengths of the directed edges $(i, j) \in E$. The graph $\tilde{G} = (\tilde{N}, \tilde{E})$ is called subgraph of $G$ if $\tilde{N} \subseteq N$ and $\tilde{E} \subseteq E$ such that all starting/end nodes of arcs of $\tilde{E}$ belong to $\tilde{N}$. Two types of subgraphs are of interest in the following: For $N^\star \subset N$ the graph $G[N^\star]$ is called $N^\star$-node-induced subgraph of $G$ if $N(G[N^\star]) = N^\star$ and $E(G[N^\star]) \subseteq E$ is the set of directed edges of $G$ connecting the nodes of $N^\star$. The second subgraph type is a walk from $n_1 \in N$ to $n_2 \in N$, denoted by $W_{n_1 n_2}$. Here, $E(W_{n_1 n_2}) = \{e_1, \ldots, e_s, \ldots, e_m\}$ is a sequence of directed edges with $e_1 = (n_1, j_1)$, $e_s = (j_{s-1}, j_s)$, $s \in \{2, \ldots m-1\}$, $e_m = (j_{m-1}, n_2)$ and $N(W_{n_1 n_2}) = \{n_1, j_1, j_2, \ldots, j_{m-1}, n_2\}$. $m$ describes the number of directed edges and

$$v(W_{n_1 n_2}) = \sum_{(i,j) \in E(W_{n_1 n_2})} v_{ij}$$

the value of $W_{n_1 n_2}$. $W^\star_{n_1 n_2}$ with $v(W^\star_{n_1 n_2}) = \min\, v(W_{n_1 n_2})$ is called minimum (valued) walk among all walks from $n_1$ to $n_2$. The determination of minimum walks is known for long (see, e.g., Dijkstra 1959) and will be used in Sect. 3. For $i \in N$ the elements of the sets $N^+(i) = \{j \in N | (i, j) \in E\}$ or $N^-(i) = \{j \in N | (j, i) \in E\}$ are called successor or predecessor nodes of $i$. With $|M|$ as notation for the cardinality of a set $M$ the number $d^+(i) = |E^+(i)|$ of outgoing edges from $i$ (with $E^+(i) = \{e \in E | e = (i, j), j \in N^+(i)\}$) and the number $d^-(i) = |E^-(i)|$ of incoming edges to $i$ (with $E^-(i) = \{e \in E | e = (j, i), j \in N^-(i)\}$) will be needed. $d^+(i)$ (or $d^-(i)$) is called out-degree (or in-degree) of $i \in N$.

Since information presented by a Web page can normally be assigned to different topic categories $c_k, k = 1, \ldots K$, denote by $(cont(i))^T = (\ldots, cont(i, c_k), \ldots)$ the content distribution for page $i$ with respect to interesting topic categories. With $Pr$ as notation for probability, $\alpha$ as probability of following the link structure of the Web pages respectively $1 - \alpha$ as probability of jumping to an arbitrary page, and $\beta$ as probability of showing interest in the same topic category respectively $1 - \beta$ as probability of shifting interest to a different topic category, a straightforward model for the navigational behavior of a random surfer can be described as follows: If $(j, c_l)$ denotes the event that a surfer visits page $j$ and shows interest in topic category $c_l$ the transition $(j, c_l) \longrightarrow (i, c_k)$ is of interest for which the following cases have to be distinguished. Given $(j, c_l)$ either (s)he will follow the link structure to page $i \in N^+(j)$ and shows interest in the same topic category $k = l$ (search strategy: $F_S$ (follow link structure, same category)) or (s)he will follow the link structure to page $i \in N^+(j)$ but now shifts interest to a different category $k \neq l$ (search stategy: $F_D$ (follow link structure, different category)), or (s)he will jump to another node $i \notin N^+(j)$ (search strategy: $J$ (jump)). Then, the importance of a Web page can be given by the probabilities

$$Pr((i,c_k)) = \sum_{j \in N^-(i)} Pr((i,c_k)|(j,c_k),F_S)\, Pr(F_S|(j,c_k))\, Pr((j,c_k))$$
$$+ \sum_{j \in N^-(i)} \sum_{\substack{l=1 \\ l \neq k}}^{K} Pr((i,c_k)|(j,c_l),F_D)\, Pr(F_D|(j,c_l))\, Pr((j,c_l))$$
$$+ \sum_{j \notin N^-(i)} \sum_{l=1}^{K} Pr((i,c_k)|(j,c_l),J)\, Pr(J|(j,c_l))\, Pr((j,c_l)) \quad (1)$$

where for the conditional probabilities it can be assumed that they have the form

$Pr(F_S|(j,c_k)) = \alpha \cdot \beta$, $Pr(F_D|(j,c_l)) = \alpha \cdot (1-\beta)$
$Pr(J|(j,c_l)) = 1-\alpha$, $Pr((i,c_k)|(j,c_k),F_S) = \frac{1}{d^+(j)}$
$Pr((i,c_k)|(j,c_l),F_D) = \frac{cont(i,c_k)}{d^+(j)}$, $Pr((i,c_k)|(j,c_l),J) = \frac{cont(i,c_k)}{n}$

with $n = |N|$. $(d^+(j))^{-1}$ can be used to describe the probability of following a link from j to another node of $N^+(j)$ (the larger $d^+(j)$ the less the importance of an outgoing link from $j$). After simplifications of (1) one gets what has been called "Topical Link Analysis" (Nie et al. 2006). If details concerning the topic categories are not known one can use $(cont(i))^T = (\ldots, \frac{1}{K}, \ldots)$ for the content distribution of the pages. In that case, the topical link analysis based approach reduces to the original PageRank model (Page et al. 1999) as explained in the next section.

### 2.2 Desirable features of Web page importance rankings and related work

There are so many reasons why Web users browse the internet that one single approach to compute a page importance ranking is not able to consider all possible requirements. Desirable features that rankings of Web pages should take into account can be described by keywords as, e.g., personalization, query dependence and topic sensitivity, dynamic nature of Web information, and offline/online computation. Short explanations for these features could run as follows: *Personalization:* Web users have different interests. Thus, algorithms for Web page importance rankings should allow for different solutions according to personal preferences. *Query dependence and topic sensitivity:* Web users want answers with respect to queries that they have formulated to get support from search engines. Here, algorithms should incorporate information about query-related topic categories which can help to find good solutions. *Dynamic nature of Web information:* The link structure of Web pages is time-dependent as new pages are created, no longer up-to-date pages are removed, and still existing pages undergo refreshments. Thus, algorithms should permanently recheck (parts of) the internet with respect to content and structure. Here, Web decomposition strategies concerning which parts of the Web should be rechecked how often are helpful. *Offline/online computations:* Online algorithms compute solutions while crawling the Web. A page ranking technique can be classified as offline algorithm if essential information (e.g., the link structure of the Web) has to be provided offline before computations can start.

In order to develop Web page importance ranking algorithms that take features of the just mentioned kind into account adequate data and computational devices have to be at hand. Additional information such as, e.g., how often and how long Web pages are looked at by which user segments, can be collected by search engine operators and added to the ranking computations.

An early approach to Web page ranking is called PageRank (Page et al. 1999). The PageRank ranking value $r_{PR}(i)$ of page $i$ is given by

$$r_{PR}(i) = \alpha \underbrace{\sum_{j \in N^-(i)} \frac{r_{PR}(j)}{d^+(j)}}_{\text{Follow Link Structure}} + (1-\alpha) \underbrace{\frac{1}{n}}_{\text{Jump to Node}} \quad , \quad i \in N, n = |N| \tag{2}$$

with $\alpha$ as probability of following the link structure of the Web pages respectively $1-\alpha$ as probability of jumping to an arbitrary node (as already mentioned in the description of the topical link analysis based approach in the last section). PageRank is neither personalized, nor query-dependent or topic sensitive, does not consider the dynamic nature of Web information, and performs offline because the link structure of the Web has to be known as the out-degrees $d^+(j)$, $j \in N^-(i)$, for the predecessors of page $i$ are needed for calculations. The mathematical explanation of (2) is based on the computation of the stationary distribution of a Markov chain that describes the random navigational behavior of Web surfers but as the Web link graph (the graph depicting the transitions between the states of the random variables of the Markov chain which are described by the Web pages) is not strongly connected, either an additional virtual node (that is linked to all other Web pages and to which all other Web pages have a link) is needed or a second graph (which models all jump possibilities between Web pages) has to be superimposed on the underlying Web link graph. A formal derivation will not be given as the PageRank model is so well-known, but having been the starting point for many modifications/generalizations it has to be mentioned in the beginning. Often, the power method is used for the computation of the PageRank ranking values (as the calculation of the stationary distribution corresponds to an eigenvalue problem) and even Krylov subspace methods (Gleich et al. 2004) have been applied to solve the underlying linear system.

The topical link analysis based approach of the last section (Nie et al. 2006) enlarges the PageRank model by additionally taking into account topic categories $c_k, k = 1, \ldots, K$, and content distributions $(cont(i))^T = (\ldots, cont(i, c_k), \ldots)$ for Web pages $i \in N$ with respect to the topic categories. Thus, this approach is topic sensitive for which, however, the provision of the content distributions is the main burden.

Attempts to consider query dependence have also been described. Let $q$ denote a query and $imp(c_k, q)$ the importance of topic category $c_k$ for query $q$. Sets $N_k \subseteq N$ of Web pages that are especially important for topic category $c_k$ can be preselected and $c_k$-dependent PageRank rankings $r^k_{PR}$ can be computed with respect to the

$N_k$-node-induced subgraph $G[N_k]$, i.e.,

$$r_{PR}^k(i) = \alpha \sum_{j \in N_k^-(i)} \frac{r_{PR}^k(j)}{d^+(j)} + (1-\alpha) \begin{cases} \frac{1}{n_k}, & i \in N_k, n_k = |N_k| \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$k = 1, \ldots, K$.

In that case a composite query-dependent page ranking $r_{PR}^{COM}$ based on precalculated $c_k$-dependent PageRanks is

$$r_{PR}^{COM}(i) = \sum_{k=1}^{K} r_{PR}^k(i)\ imp(c_k, q)$$

as suggested by Haveliwala (2002). Here, problems arise with respect to the selection of the graphs $G[N_k]$ and the determination of $imp(c_k, q)$ values for the different queries that Web users may formulate.

Even more ambitious was the suggestion by Richardson and Domingos (2002). With $imp(i, q)$ as importance of page $i$ for query $q$ and $\frac{imp(i,q)}{\sum_{j \in N} imp(j,q)}$ as query-dependent selection probability for page $i$ one gets query-dependent ranking values

$$r_{PR}^q(i) = \alpha \sum_{j \in N^-(i)} r_{PR}^q(j) \frac{imp(i,q)}{\sum_{l \in N^+(j)} imp(l,q)} + (1-\alpha) \frac{imp(i,q)}{\sum_{j \in N} imp(j,q)}, i \in N \quad (4)$$

which could be computed by just replacing the probabilities $\frac{1}{d^+(j)}$ and $\frac{1}{n}$ in the PageRank formula (2) by query-dependent counterparts. Unfortunately, the problem of how $imp(i, q)$ values could be determined immediately after a query has been formulated poses a serious challenge.

A similar idea as used in approach (4) has been proposed to personalize the original PageRank by substituting

$$\frac{pref(i)}{\sum_{l \in N^+(j)} pref(l)} \quad \text{for} \quad \frac{1}{d^+(j)} \quad \text{and} \quad \frac{pref(i)}{\sum_{l \in N} pref(l)} \quad \text{for} \quad \frac{1}{n}$$

where $pref(i)$ describes the (personal) preference for page $i$ (think of situations where Web users are invited to provide judgements with respect to books, hotel accomodation, traveling conditions, and the like).

In the HITS (Hyperlink Included Topic Search) approach (Kleinberg 1999) one starts with a properly selected subset $N^*$ (based on a query $q$ or important topic categories) and restricts ranking calculations to $G[N^*]$ with adjacency matrix $A^*$. Additionally, it is argued that each page can be characterized by two rankings as authority (A good authority is a page that is linked to by many good hubs.) as well as hub (A good hub is a page that has links to many good authorities.) and, consequently,

the underlying situation is described by an authority ranking $r_a$ and a hub ranking $r_h$. With

$$r_a = A^{*T} r_h \quad \text{and} \quad r_h = A^* r_a \quad \text{one gets} \quad r_a = A^{*T} A^* r_a \quad \text{and} \quad r_h = A^* A^{*T} r_h$$

and can solve the problem via power method application in a considerably reduced subgraph $G[N^*]$. The formulation of a topical link analysis based version, called Topical HITS, was already described in Nie et al. (2006).

Of course, the recognition of the increasing interest in Web page importance rankings has led to attempts to manipulate the positionings in ranking lists. On the other side, developers of page ranking approaches have tried to secure their techniques against undesirable influences. An example is the TKC (Tightly Knit Community) effect which gives pages within a small but highly interconnected set of pages a high ranking even though the pages are not relevant or pertain to just one aspect of a topic.

Thus, Lempel and Moran (2000) have suggested SALSA (Stochastic Approach for Link Structure Analysis) which modifies the HITS approach by transfering the link structure of $G[N^*]$ to the interconnections between a hub side and an authority side of an undirected bipartite graph which are used for ranking calculations that are less influenced by the TKC effect.

Finally, OPIC (Online Page Importance Computation) by Abiteboul et al. (2003) should be mentioned as an early possibility to compute page rankings online and to take the dynamic nature of Web information into account. Here, the Web is crawled and in the t-th crawling step two values, $\text{cash}(i,t)$ and $\text{hist}(i,t)$ are assigned to each page $i \in N$ (with $\sum_{i \in N} \text{cash}(i,0) = 1$ and $\text{hist}(i,0) = 0, i \in N$) and updated by distributing the cash of each page $j$ just visited equally to its successors (more precisely: $\text{cash}(j,t+1) = 0$, and $\text{hist}(i,t+1) = \text{hist}(i,t) + \frac{\text{cash}(j,t)}{d^+(j)}, i \in N^+(j)$, all other values remain unchanged). The idea is that cash normalized by cumulated history (hist) values (where hist of a page sums the cash received by predecessor nodes) can be used for page ranking.

The authors argue that the vector with the components $\frac{\text{cash}(i,t)+\text{hist}(i,t)}{(\sum_{i \in N} \text{hist}(i,t))+1}$ converges for $t \to \infty$ to a page importance ranking of the Web. Different crawling strategies (cyclic, greedy, random) and windowing techniques (how data are gathered in intervalls $[t-T,t]$) influence the calculations. OPIC is link structure based, can be performed online, and is able to consider dynamic and category-related aspects by appropriate assignments of cash.

## 3 m-Clicks-ahead recommendations

Although further references concerning page rankings could have been mentioned in the preceding section, the background discussion has provided enough essential aspects, e.g., that the development of Web page importance rankings started with considerations to model the random navigational behavior of Web surfers in underlying link graphs and that soon attempts to incorporate additional data $z_i = (\ldots, z_{ix}(x = 1, \ldots, X), \ldots)$ for the description of pages (e.g., importance

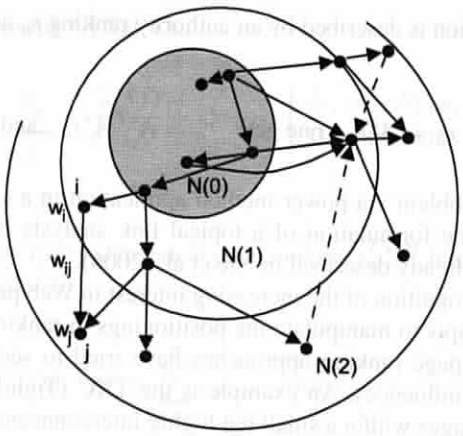


**Fig. 1** m= 2-clicks-ahead situation

with respect to topic categories) and $z_{ij} = (\ldots z_{ijy}(y = 1, \ldots, Y), \ldots)$ for pairs of pages (e.g., (dis)similarities with respect to content) were formulated in order to overcome shortcomings of the early ranking suggestions. Nowadays, often (overly) long lists of relevant Web pages are presented as solution of the ranking process for which the way how these lists were determined is not revealed. For a Web user, however, a complete ranking of all possible pages is of less importance and empirical tests have shown that essentially Web addresses from the first positions in corresponding lists are selected for information collection, hence, the following situation is of interest:

Given that a Web user has collected some Web pages that (s)he finds suitable, an obvious possibility to look for additional relevant information is to try a few links provided by the just chosen pages in order to obtain further interesting information. Here, the link structure of eligible pages has to be investigated, for which purpose a m-clicks-ahead search strategy can be implemented by means of crawling and supported in terms of minimum walk methodology.

### 3.1 Minimum walk based m-clicks-ahead rankings

Let $N(0)$ denote the set of personally interesting Web pages and $N(m)$ the set of pages that can be reached from $N(0)$ via at most $m$ links (clicks) from a node of $N(0)$. The link structure description of this m-clicks-ahead situation is depicted in Fig. 1.

As all outgoing arcs from nodes of $N(m)$ are given as corresponding links of the underlying pages, one has for the successor nodes $N^+(i, m) = N^+(i), i \in N(m)$, for increasing $m$ but for the predecessors $N^-(i, m)$ of nodes of $N(m)$ in the m-clicks-ahead situation only $N^-(i, m) \subseteq N^-(i), i \in N(m)$, and the condition

$$N^-(i, m') \supseteq N^-(i, m), \quad m' > m, \quad i \in N(m) \tag{5}$$

is valid as there might be links from pages of $N(m')$ to pages of $N(m)$ that are still not known in the m-clicks-ahead situation. From

$$v_{ij}(m) = f(\ldots, z_i(m), \ldots, z_j(m), \ldots, z_{ij}(m), \ldots) \tag{6}$$

with $f$ as function that describes a cost term for link $(i, j)$ and $v_i(m)$ as value of a minimum walk from a node of $N(0)$ to node $i$ of $N(m)$ one gets as importance ranking

$$r(v_i(m)) = \frac{M(m) - v_i(m)}{|N(m)| \cdot M(m) - S(m)}, \quad i \in N(m) \tag{7}$$

with $M(m) = \max_{i \in N(m)} v_i(m)$ and $S(m) = \sum_{i \in N(m)} v_i(m)$.
$z_i(m)$ stands for page $i$ information based on $N(m)$ and $z_{ij}(m)$ for data with respect to pairs of pages related via an arc $e = (i, j) \in E(m)$ as set of links that can be provided after $m$ clicks. For different $m$ the minimum walk recommendations based on $v_i(m), i \in N(m) \setminus N(0)$, can change as condition (5) shows. Furthermore, for increasing $m$ the number of pages not directly selected because of personal interest but provided by the crawling process can increase drastically (in other words: the percentage $\frac{|N(0)|}{|N(m)|}$ of initially selected pages in relation to $N(m)$ can become very small) so that the content of pages from $N(m) \setminus N(0)$ and their outgoing links may overrule the intended aim of finding additionally interesting pages that are highly related to the pages of $N(0)$. Thus, only $m \in \{1, 2\}$ will be used in the following. Additionally, the manner in which valuation (6) is implemented has to be discussed. As $v_{ij}(m)$ describes a cost-term one gets

| | |
|---|---|
| the larger $d^+(i)$ | the larger $v_{ij}(m)$ |
| the larger $d^-(j)$ | the smaller $v_{ij}(m)$ |
| the larger $imp(i, c_k)$ | the smaller $v_{ij}(m)$ |
| (and/or $imp(j, c_k)$) | (if category-dependent) |
| $\vdots$ | $\vdots$ |

if, e.g., $d^+(i), imp(i, c_k), \ldots$ are components of $z_i(m)$. Similar considerations apply to the components of $z_{ij}(m)$. Hence, formula (6) is very flexible in incorporating additional data that may help to better describe the relationships between pages in terms of personalization, query dependence, topic sensitivity and the like. As crawling is used, the dynamic nature of Web information is taken into consideration and online computation becomes possible. In first applications

$$v_{ij}(m) = \begin{cases} \log(d^+(i) + \epsilon) \exp(-d^-(j, m) \cdot c^{-1}), & (i, j) \in E(m) \\ C, & \text{otherwise,} \end{cases} \tag{8}$$

was used as an example for a valuation based on link structure information with respect to $E(m)$. Here, $\epsilon > 0$ is a factor to ensure $v_{ij}(m) > 0$, $d^-(j, m)$ is a lower bound for the in-degree $d^-(j)$ of page $j$ restricted to $E(m)$, and $c, C$ are sufficiently large constants. A crawling based approach with valuation $v_{ij} = \log(d^+(i))$ (Bidoki et al. 2007) turns out to be a special case of (8). For the crawling process, restrictions

that exclude inadmissable or undesirable links (e.g., links that point to parts of the underlying page, links that point to ads) have to be installed.

*3.1.1 Content similarity adjustments*

Link structure based page rankings use the fact that Web pages contain links to other pages that are (claimed to be) relevant to the content of the actually visited page. The ideal situation would be that only links to pages with information really important for the meaning of the underlying page are added but in reality links can exist that point to Web pages (remember the TKC effect mentioned earlier) that a surfer would not have visited, normally. A possibility to lower the influence of less relevant or overly dominant (because of extraordinary high in-degree values) pages within m-clicks-ahead search strategies consists in an additional check for similarity of the contents of the pages of $N(0)$ and the contents of solutions returned within the linkage based computations.

As starting point a vocabulary $V$ is needed as set of salient words or terms to desribe the phenomenon for which the pages of $N(0)$ give first answers. With $V = V(N(0))$ a Web page $i$ can be viewed as a document, in which the appearance of terms of $V$ is taken into consideration and represented by a page document vector $w_i \in \mathbb{R}^{|V|}$ whose components are adequately weighted frequencies with which the terms of $V$ appear in page $i$. With $\bar{w}$ as mean vector of the page document vectors of $N(0)$

$$dis_i = dis(\bar{w}, w_i) \in [0, 1], \quad i \in N(m) \tag{9}$$

can be calculated (e.g, with the cosine coefficient as measure for vector similarity) to indicate how dissimilar the content of page $i$ is with respect to the contents of the starting pages of $N(0)$. Then, a content similarity adjusted minimum walk value

$$\tilde{v}_i(m) = g(v_i(m), dis_i), \quad i \in N(m) \tag{10}$$

(with $g$ as function that combines the already known $v_i(m)$ values with content dissimilarity judgements of page $i$ relative to $N(0)$) allows to re-rank the Web pages in the m-clicks-ahead neighbourhood of $N(0)$. In first applications

$$\tilde{v}_i(m) = \frac{v_i(m)}{M(m)} + \frac{1-\gamma}{\gamma} dis_i, \quad \gamma \in (0, 1] \tag{11}$$

was used. For $\gamma = 1$ one gets the importance ranking as described by formula (7), for $\gamma \to 0$ content similarity becomes more and more important. Notice that link structure based information is essential for the calculation of the m-clicks-ahead minimum walk strategies and content similarity checks aim at hindering pages that are less important for the situation under consideration to receive high ranking positionings.

*3.1.2 Acceleration of ranking computations*

The ranking computations according to formula (7) are dependent on $N(m)$ and the size of $N(m)$ and in applications $|N(m)|$ can become large. Hence, an acceleration

**Table 1** First 10 recommendations of the linkage structure based variant ($\gamma = 1$) of the m-clicks-ahead search (cold aspects)

| $m = 1$ ($\gamma = 1$) | $m = 2$ ($\gamma = 1$) |
|---|---|
| N(0){ Schweinegrippe, Grippeimpfung, Fieber | N(0){ Schweinegrippe, Grippeimpfung, Fieber |
| 1. Schweineinfluenza | 1. Schweineinfluenza |
| 2. Pandemie_H1N1_2009/10 | 2. Pandemie_H1N1_2009/10 |
| 3. Immunsystem | 3. Latein |
| 4. Entz%C3%BCndung | 4. Immunsystem |
| 5. Latein | 5. Weltgesundheitsorganisation |
| 6. Antigen | 6. Altgriechische_Sprache |
| 7. Influenza | 7. Mutation |
| 8. Impfung | 8. Antigen |
| 9. Antibiotika | 9. Antibiotika |
| 10. Lungenentz%C3%BCndung | 10. Impfung |

to provide first rankings at "earlier" crawling steps and replace them when additonal informations are available is an option. If $p_l, l = 1, \ldots, L$, are the first Web pages for which $N^+(p_l)$ has already been checked – dependent on the crawler implementation – it might be of interest to use the set of successors $N^+(p_1, \ldots, p_L)$ of $\{p_1, \ldots, p_L\}$ for rapid ranking calculations (Note that if $\{p_1, \ldots, p_L\} = N(0)$ one gets $N^+(p_1, \ldots, p_L) = N(1)$.). The idea is not to wait until $v_i(m)$ values are available for all pages of $N(m)$ but to apply a properly modified formula (7) restricted to $N^+(p_1, \ldots, p_L)$ and to update the rankings for increasing $L$.

### 3.2 Examples

The following examples show what can be expected from an application of a m-clicks-ahead search strategy.

*Example 1* Assume that a Web user is interested in cold aspects, has visited Wikipedia (German version), and selected the pages http://de.wikipedia.org/wiki/Schweinegrippe (swine flu) and http://de.wikipedia.org/wiki/Grippeimpfung (flu shot) together with page http://de.wikipedia.org/wiki/Fieber (fever) as starting information. Instead of personally checking all the links of these pages, the m-clicks-ahead search crawls $|N(1)| = 235$ and $|N(2)| = 10{,}251$ pages and gives recommendations as depicted in Table 1 within the linkage structure based variant ($\gamma$=1) of the algorithm. The starting part http://de.wikipedia.org/wiki/ of the URLs has been omitted and representation is restricted to the top 10 recommendations (as the 10 positions in Google's first page of the complete list of entries with respect to an underlying query).

The recommendations to look for Schweineinfluenza and Pandemie_H1N1_2009/10 remain stable in both rankings of Table 1 while, e.g., the positioning of Immunsystem worsens and the positioning of Weltgesundheitsorganisation (which was on ranking position 11 for $m = 1$) improves with respect to $m = 2$.

Links such as Weltgesundheitsorganisation or (even more adequate to demonstrate the already mentioned phenomenon) Altgriechische_Sprache gain importance for

**Table 2** First 10 recommendations of the content similarity adjusted variant ($\gamma = 0,6$) of the 2-clicks-ahead search (cold aspects)

| $m = 2$ ($\gamma = 0,6$) |
|---|
| N(0) { Schweinegrippe, Grippeimpfung, Fieber } |
| 1. Pandemie_H1N1_2009/10 |
| 2. Schweineinfluenza |
| 3. Latein |
| 4. Impfung |
| 5. Immunsystem |
| 6. Influenza |
| 7. Pandemie_H1N1_2009 |
| 8. Mutation |
| 9. Ethik |
| 10. Tuberkulose |

increasing $m$ as they are of general interest and more and more pages from $N(m)\backslash N(0)$ have a link to this kind of information.

In a pure content similarity based computation of the 2-clicks-ahead search strategy, e.g., Altgriechische_Sprache has positioning 3,548 and Weltgesundheitsorganisation has positioning 8,089 within the list of 10,251 pages mentioned. Table 2 shows the 10 highest ranked positionings of the content similarity adjusted 2-clicks-ahead solution for $\gamma = 0,6$.

The choice of the value of $\gamma$ depends on $N(0)$ (and the vocabulary $V(N(0))$). For $\gamma > 0,5$ the linkage structure of the Web dominates the ranking calculations. A good mix between linkage structure and content similarity information with respect to the underlying example is yielded for $0.4 \leq \gamma \leq 0.6$.

*Example 2* Assume that a Web user wants to know some details about the worldchampionship in football/soccer in South Africa 2010 and has selected http://www.southafrika2010.de, http://www.fd21.de/439515.asp, and http://de.fifa.com/worldcup as starting set $N(0)$ (end of September 2010). $|N(1)| = 169$ and $|N(2)| = 2{,}187$ pages were crawled. An earlier application with these starting pages (end of July 2010) resulted in larger $N(1)$, $N(2)$ sets.[1]

Table 3 shows the top 10 positionings within the linkage structure based variant ($\gamma = 1$) of the m-clicks-ahead methodology.

The six permutations of the keywords southafrika2010, worldcup, and fifa used as queries yielded between about 11,700,000 and 30,700,000 entries in Google lists. A comparison of the top positionings in these lists with m-clicks-ahead recommendations will not be made since the way how Google rankings are obtained is not known. In Table 3 the importance of the southafrika2010 site diminishes for $m = 2$ because 'after the last championship' is 'before the next championships', which demonstrates that more up–to-date information with respect to future events such as, e.g., em-2012, wm-2014, em-2016 (em $\cong$ European football championship, wm $\cong$ World football

[1] A first version of this m-clicks-ahead search strategy was presented at the Joint Meeting of GfKl-CLADAG 2010, September 8–10, 2010, Florence, Italy.

**Table 3** First 10 recommendations of the linkage based variant ($\gamma = 1$) of the m-clicks-ahead search (football world championship)

| $m = 1$ ($\gamma = 1$) | $m = 2$ ($\gamma = 1$) |
|---|---|
| N(0) { www.southafrika2010.de<br>www.fd21.de/439515.asp<br>de.fifa.com/worldcup | N(0) { www.southafrika2010.de<br>www.fd21.de/439515.asp<br>de.fifa.com/worldcup |
| 1. southafrika2010.de/suedafrika/aktuelle-fifa-weltrangliste | 1. fussball-em-2012.com |
| 2. southafrika2010.de/fifa-weltrangliste-juni-2010 | 2. fussball-wm-2014.biz |
| 3. southafrika2010.de/suedafrika/bester-spieler-der-wm-diego-forlan | 3. bloggerei.de/rubrik_3_Sportblogs |
| 4. southafrika2010.de/suedafrika/ruecken nummern-der-deutschen-nationalmannschaft | 4. topblogs.de |
| 5. fussball-em-2012.com | 5. fussball-em-2016.com |
| 6. southafrika2010.de/feed | 6. southafrika2010.de/wm-2010 |
| 7. southafrika2010.de/suedafrika/fuenf-baustellen-der-dfb-nationalmannschaft | 7. southafrika2010.de/wm-2010-tabelle/wm-2010-viertelfinale |
| 8. southafrika2010.de/suedafrika/jogi-loew-verlaengert-bis-zur-em-2012 | 8. southafrika2010.de/wm-2010-tabelle/wm-2010-halbfinale-finale |
| 9. southafrika2010.de/suedafrika/dfb-team-entschuldigt-sich-bei-den-fans | 9. southafrika2010.de/wm-2010-trikotshop |
| 10. southafrika2010.de/suedafrika/fussball-weltrangliste-juli-2010 | 10. southafrika2010.de/wm-2010-tabelle/wm-2010-achtelfinale |

**Table 4** First 10 recommendations of the content similarity adjusted variant ($\gamma = 0,8$) of the 2-clicks-ahead search (football world championship)

| $m = 2$ ($\gamma = 0,8$) |
|---|
| N(0) { www.southafrika2010.de<br>www.fd21.de/439515.asp<br>de.fifa.com/worldcup |
| 1. southafrika2010.de/wm-2010 |
| 2. fussball-em-2012.com |
| 3. southafrika2010.de/wm-2010-trikotshop |
| 4. southafrika2010.de/wm-2010-tabelle/wm-2010-halbfinale-finale |
| 5. southafrika2010.de/wm-2010-tabelle/wm-2010-achtelfinale |
| 6. southafrika2010.de/suedafrika/aktuelle-fifa-weltrangliste |
| 7. southafrika2010.de/wm-2010-tabelle/wm-2010-viertelfinale |
| 8. southafrika2010.de/suedafrika/fussball-weltrangliste-juli-2010 |
| 9. southafrika2010.de/suedafrika/fuenf-baustellen-der-dfb-nationalmannschaft |
| 10. southafrika2010.de/suedafrika/dfb-team-entschuldigt-sich-bei-den-fans |

championship) and URLs for blogs are gaining importance in the linkage structure based evaluation.

Again, variants of the content similarity adjusted m-clicks-ahead search strategies were checked. Table 4 shows the 10 highest ranked positionings of the content similarity adjusted 2-clicks-ahead solution for $\gamma = 0.8$. Already for $\gamma = 0,8$ the URLs describing blogs and some of the em-, wm-sites of the recommendations for

$m = 2$ ($\gamma = 1$) of Table 3 are no longer in the list of the top ranked Web addresses. This indicates that linkage structure based methods can be essential if one wants to uncover newly created contents that give hints to future developments (as content similarity between descriptions of present and future situations based on terms used by the page document vectors of $N(0)$ may be low) or aspects not in the main focus of $V(N(0))$.

The just described examples were selected in order to impart a feeling of how the m-clicks-ahead search strategy works. Of course, solutions can depend on data (additional to link structure information and content similarity) as collected by search engine operators and on the manner in which crawling was implemented, on the starting set $N(0)$ (e.g., whether solely Wikipedia URLs or pages from different Web sources are used and the size of $N(0)$), and on the importance of pages from $N(m) \setminus N(0)$ (e.g., the dominance of a site within $N(m) \setminus N(0)$, the increasing actuality of new pages, the general significance of certain pages for other pages)—to mention just some salient aspects. Here, a Web user can select different starting sets $N(0)$, vary the number $m$ within m-clicks-ahead search strategies, and the value of $\gamma$ in order to find trends and/or underpin her/his opinion.

## 4 Concluding remarks

In order to optimize search strategies to find Web pages that are relevant for corresponding purposes considerable efforts have been made to rank pages according to their importance with respect to desirable features. While early approaches were mainly based on link structure considerations, it was soon realized that additional data and suited support tools would be needed for more realistic Web page importance rankings: For a single page, e.g., data that describe content and suitability of the page with respect to interesting topic categories as well as queries raised by Web users; for a pair of pages, e.g., data that account for their (dis)similarity in relation to sets of category or query dependent starting pages. To avoid a lengthy repetition of explanations provided in earlier parts of this article, Table 5 depicts a collection of selected data types that were mentioned in preceding descriptions and shows that there are challenges for new directions within data analysis methodology (see, e.g., Gaul 2006 for further data analysis applications to Web mining).

Nowadays, search engine operators do not reveal how Web page importance rankings are determined, but data as listed in Table 5 could be helpful. Some ranking activities to provide data for the consideration of desired features can be performed in advance, e.g., for frequently used queries (sets of) query-related pages can be preselected, for collections of pages relating to important topic categories category-dependent rankings can be precalculated, for optimized assignments between important topic categories and frequently used queries computations can be done in advance, and (dis)similarities between the contents of pages can be checked by borrowing ideas from text mining.

Additionally to the page ranking literature cited in earlier parts of this article introductions into network analysis (e.g., Brandes and Erlebach 2005), the science of search

**Table 5** Selected data for Web page importance rankings

| | |
|---|---|
| $A = (a_{ij})$ | Adjacency matrix of underlying Web link graph |
| $d^+(i), d^-(i)$ | Out-degree, in-degree of page $i$ |
| $d^-(i, m)$ | In-degree of page $i$ restricted to $N(m)$ |
| $\vdots$ | — Link structure related data — |
| $imp(i, c_k)$ | Importance of page $i$ for topic category $c_k$ |
| $(\approx cont(i, c_k))$ | ($\approx$ content probability) |
| $G[N_k]$ | $N_k$-node-induced subgraph for topic category $c_k$ |
| $\vdots$ | — Category related data — |
| $imp(i, q)$ | Importance of page $i$ for query $q$ |
| $G[N_q]$ | $N_q$-node-induced subgraph for query $q$ |
| $\vdots$ | —Query related data — |
| $imp(c_k, q)$ | Importance of topic category $c_k$ for query $q$ |
| $hl(i)$ | How long page $i$ is looked at |
| $ho(i)$ | How often page $i$ is looked at |
| $\vdots$ | — Usage related data — |
| $pref(i)$ | Preference for page $i$ |
| $cash(i)$ | Cash of page $i$ } for crawling dependent online search |
| $hist(i)$ | History of page $i$ } for crawling dependent online search |
| $dis/sim(i, j)$ | (Dis)similarity between page $i$ and page $j$ |
| $v_{ij}(m)$ | Valuation of the (page $i$, page $j$)-pair within the m-clicks-ahead situation |
| $\vdots$ | — Pair of pages related data — |

engine rankings (e.g., Langville and Meyer 2006), and text processing (e.g., Salton 1989) should be mentioned for foundations as well as further directions that could be of interest for the problem addressed.

The development of the m-clicks-ahead search strategy was influenced by work on recommender systems (see, e.g., Gaul and Schmidt-Thieme 2002; Bomhardt and Gaul 2009) and applied to Web marketing (e.g., Gaul 2004). Depending on which information can be considered in $v_{ij}(m)$, reasonable choices of $N(0)$, and realistic values for $m$ and $\gamma$, useful information for the selection of additionally interesting Web pages in a m-clicks-ahead situation can be expected.

**Acknowledgments** Support by Dominique Vincent and Krastyu Georgiev with respect to the crawling process is gratefully acknowledged.

## References

Abiteboul S, Preda M, Cobena G (2003) Adaptive on-line page importance computation, WWW 2003, 280–290

Bidoki AMZ, Yazdani N, Ghodsnia P (2007) FICA: a fast intelligent crawling algorithm, IEEE/WIC/ACM international confernce on web intelligence, 635–641

Bomhardt C, Gaul W (2009) Feedback options for a personal news recommendation tool. In: Okada A et al (eds) Cooperation in classification and data analysis, Studies in classification, data analysis, and knowledge organization. Springer, pp 91—98

Brandes U, Erlebach TH (eds) (2005) Network analysis: methodological foundations, Lecture notes in computer science, vol 3418, Springer

Dijkstra EW (1959) A note on two problems in connection with graphs. Numerische Mathematik 1:269–271
Gaul W (2004) Market research and the rise of the Web: the challenge. In: Wind Y (Jerry), Green PE (eds) Market research and modeling: progress and prospects: a tribute to Paul E. Green, International series in quantitative marketing. Kluwer, pp 103–113
Gaul W (2006) Challenges concerning Web data mining, COMPSTAT 2006, pp 403–416
Gaul W, Schmidt-Thieme L (2002) Recommender systems based on user navigational behavior in the internet. Behaviometrika 29:1–22
Gleich D, Zhukov L, Berkhin P (2004) Fast parallel PageRank: a linear system approach, yahoo! research labs technical report YRL-2004-38
Haveliwala TH (2002) Topic-sensitive pagerank, Proceedings of the 11th international conference, Honolulu, Hawai
Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
Langville AN, Meyer CD (2006) Google's PageRank and beyond: the science of search engine rankings. Princeton University Press, New Jersey
Lempel R, Moran S (2000) The stochastic approach for link-structure analysis (SALSA) and the TKC effect, Proceedings of the 9th international WWW conference, Amsterdam
Nie L, Davison BD, Qi X (2006) Topical link analysis for web search, SIGIR'06, 91-98
Page L, Brin S, Motwani R, Winograd T (1999) The page rank citation ranking: bringing order to the Web, manuscript
Richardson M, Domingos P (2002) The intelligent surfer: probabilistic combination of link and content information. Adv Neural Inform Process Syst 14:1441–1448
Salton G (1989) Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley