Optimizing WebPage Interest

W.J.M. Elbers and Th.P. van der Weide

Institute for Computing and Information Sciences,
Radboud University, Nijmegen,
The Netherlands
ru@willemelbers.nl and tvdw@cs.ru.nl

Abstract. In the rapidly evolving and growing environment of the internet, web site owners aim to maximize interest for their web site. In this article we propose a model, which combines the static structure of the internet with activity based data, to compute an interest based ranking. This ranking can be used to gain more insight into the flow of users over the internet, optimize the position of a web site and improve strategic decisions and investments. The model consists of a static centrality based component and a dynamic activity based component. The components are used to create a Markov Model in order to compute a ranking.
Keywords: web graph; interest; centrality; user flow; Markov Model

1 Introduction

Users are entering the world wide web by accessing a web site and use the available hyperlinks to travel to other pages and web sites. Simultaneously web site owners are constantly updating their existing web sites and creating new web sites. Over time web sites might also cease to exist. In short, users follow the structure created by web masters and others while this structure is constantly evolving. In this article we want to investigate how we can gain more insight into the static structure of the internet and the dynamic flow of users through this structure. This results in a flow potential score for a web site. The improved insight, based on the flow potential score, can result in more strategic decisions and investments.

Flow potential, which is more than just flow if it also depends on properties of the underlying structure, will be referred to as web site interest. The research question in this article is: How can web site interest be measured based on static and dynamic properties? In order to answer this question, the following sub questions have to be answered: (1) What are the static and dynamic properties of web sites?, (2) How can these properties be measured? and (3) How can these two types of properties be combined?

In section 2 of this article the model is introduced. The static and dynamic properties will be specified in the context of an experiment, discussed in section 3. The initial results are presented in section 4 and section 5 will conclude this article.

2 The Model

The model, proposed in this section, derives a web page interest value $R(p)$ of a web page p from the following two components. The first component is the web page importance $S(p)$, which is measured relative to other web pages. The second component $D(p)$, is a property that quantifies the interest in page p. These components are combined by a function called R_c:

$$R(p) = R_c(S(p), D(p)) \,. \tag{1}$$

The two components combine static and dynamic properties of web pages respectively. The importance function S is a static property of the (web)graph and may be measured by centrality, which is a known concept from graph theory. Centrality is a measure to indicate the importance of a node in the graph, based only on the structure of the graph. The four most known centrality measures, introduced by Freeman [1] and Bonacich [2], are degree centrality, betweenness centrality, closeness centrality and eigenvector centrality. These centrality measures also have a conceptual meaning. The degree centrality measures the potential of a node to be part of the flow in a graph. Betweenness centrality can be seen as the potential of a node to control the flow in a graph. Closeness centrality can be seen as the potential of a node to avoid the control potential of other nodes in a graph. Eigenvector centrality is a measure for how connected to other influential nodes a node is in a graph. Besides these centrality measures, there are also two well known algorithms which use the static web graph to rank pages: PageRank [3] and HITS [4].

The interest function D is a flexible, dynamic, component. Link traversal counts how often users follow specific links. This would be the best activity based measure in the case of website interest. Unfortunately this information is, usually, not publicly available. Even the number of visitors of a web site is hard to obtain. We will propose a solution to convert activity based data for nodes into probabilities of following a link.

So far, the components in the model have been introduced, their relation has not. The solution is based on work in the field of adaptive web sites, [5] [6] [7] and especially [8] [9] [10] [11]. Using a Markov Model seems to be a promising solution for R_c. The m nodes of a graph are the states of the Markov Model. The, structural, centrality measure can be used to create an $1 \times m$ initial probability distribution, L, and the activity based data, which are transformed into transition probabilities, can be used as the $m \times m$ one step transition probability matrix Q. Then R will be the $1 \times m$ ranking vector $R = L \times Q^k$, based on taking k steps through the graph. At some point, for a large enough k, a steady state is reached where increasing k further has no effect anymore. That state is also independent of the initial probability distribution and at that point the ranking will only be activity based.

In the remainder of this article we will use the following definitions for graphs. A (web)graph G is defined as an ordered pair $G = (V, A)$ where $V = \{p_1, \ldots, p_n\}$ is the set of vertices or nodes and A is the set of arcs between the nodes in the

graph, defined as the set of ordered pairs $(v, w) \in A \subseteq V^2$. We can also write $v \rightarrow w$ or $A(v, w)$ to specify an arc in the graph. In the case of a web graph, the nodes of the graph are the actual web pages and the arcs are the actual hyperlinks between these web pages.

3 The Model into Action

3.1 The Static Property

Based on the static graph structure we have to compute a centrality score for each node. Four methods to compute centrality have been mentioned in section 2. Based on their conceptual meaning, betweenness centrality seems like a very promising candidate. This is a measure for the potential of a node in the graph to control the flow. If many people pass through a site, z, when following links from site v to site w, then this z has a high potential to control where those people are going.

In order to optimize the calculation for betweenness centrality, we will use ego betweenness, introduced by Everett and Borgatti [12]. Ego betweenness of a node v is the betweenness score of that node in its ego network as defined by Freeman [13]. The ego network of a node is the graph with the node itself, all the direct neighbors of this node and the arcs between these nodes in the original graph. The betweenness for each node is needed, therefore n ego networks have to be computed. The advantage of these ego networks is that they will be relatively small. We have approximately 12K, uniquely connected, nodes in the test dataset, but the average ego network size is only 10 nodes and the biggest ego network is around 250 nodes. How do we extract the ego network, $G_{ego} = (V_{ego}, A_{ego})$, for a node $v \in V$ from graph $G = (V, A)$? Based on this definition, two properties hold: (1) $V_{ego} \subseteq V$ and (2) $A_{ego} \subseteq A$ and based on the definition of Freeman all direct neighbors and their arcs of v need to be included.

We perform two steps to extract the ego network. First we will get the set with all nodes in the ego network for a certain node v: $V_{ego} = \{v\} \cup \{w \in V | (v, w) \in A \vee (w, v) \in A\}$. Second, based on the set with nodes in the ego network, V_{ego}, we can construct the set of arcs in the ego network, A_{ego}. If an arc $(v, w) \in A_{ego}$, exists in A then it should also exist in A_{ego}: $A_{ego} = \{(v, w) \in A | v \in V_{ego} \wedge w \in V_{ego}\}$.

Now that we have the ego network for a node v in place, its actual ego betweenness score, c_b, can be computed. Let $B_{ego} = A_{ego}^2 \times (1 - A_{ego})$ where 1 is a matrix with only ones of the same dimension as A_{ego} and $\times$ is the cellwise multiplication operator for matrices. The ego betweenness is the sum of the reciprocals for the non zero entries in B_{ego}:

$$c_b = 1/ \parallel B_{ego} \parallel_1 \ . \tag{2}$$

If the ego betweenness is computed for all nodes in the graph, the result will be a vector C_b with these scores for each node. Next, this vector is transformed into

the initial probability distribution by dividing each centrality score by the sum of all centrality scores:

$$I = 1/ \parallel C_b \parallel_1 \times C_b \,. \tag{3}$$

Degree centrality could also be an interesting measure to use. It is the potential of being part of a flow in the graph. However, if you are not part of any shortest paths between two web sites, people are more likely to follow the shorter paths and not enter your web site. Degree centrality is an easy to compute centrality measure, therefore it might be interesting to compare degree centrality based rankings to betweenness centrality based rankings.

Closeness centrality is the potential of a node to avoid being part of the flow. Since we are interested in optimizing the flow to our own web site, we are not so much interested in web sites which can avoid the flow of other web sites. This could be a desirable measure if information independence is very important.

Eigenvector centrality is a measure which increases a nodes importance if it is connected to other important nodes. Since this centrality measure is less aimed at how a node can influence the flow in a graph, we didn't choose to use this centrality measure. However, after the first experiments, it could be interesting to see how this centrality measure fits in and performs.

PageRank would also be a interesting measure to use for the static property. To put it simple, a high PageRank is an indication for the number of incoming pages and their PageRank. Therefore it seems quite likely to say there should be some relation between a high PageRank and a high flow, however many incoming links do not necessarily mean a lot of incoming traffic. We think this is the most interesting alternative to look into for any future research. The HITS algorithm assigns hub and authority scores to the nodes in the graph. It is not obvious how this relates to the flow in a graph, since the number of links doesn't say anything about traffic numbers directly. The chance on more traffic might be bigger with more incoming or outgoing links, but this requires further research.

3.2 The Dynamic Property

As mentioned already, it would be ideal to have link traversal or traffic data of all web sites on the internet. Unfortunately this is not possible. We have come up with a different approach to work around this problem. This approach is applicable to websites as well as blogs, as long as usage data is available for the node in the graph. Because we have a dataset of the dutch blogosphere, we have come up with a solution based on timestamps as a measure for blog activity. This method can also be used for fora, but for websites a different approach has to be used in order to retrieve the activity based data. The basic concept of converting node based activity into link traversal activity, as proposed in the following sections, is also applicable to websites as a whole instead of blogs only.

Since the dataset contains blogs, we will propose a method to crawl activity based data from blogs. Blogs often have the option to post reactions with a topic. These reactions can be characterized by a time stamp on the page. For our experiment we will gather all time stamps associated with a blog and use

this as the activity measure for the dynamic property. This approach is based on the assumption that reactions to a blog are related with the traffic of that blog. This approach has a big advantage, it's easy to add into the crawling process which analyzes the blogs to construct the graph structure. By using these time stamps as a measure for the number of reactions on a blog, we have an easy way to obtain a measure for the activity on a blog.

Let $G = (V, A)$ be a graph consisting of a set of nodes V and a set of arcs, $A \subseteq V^2$. For each node $v \in V$ the function $r(v)$ returns the activity measure for the supplied node v. In our case this activity measure is the number of reactions that were posted on a blog. Blog visitors are traveling this network structure. Let $P(w|v)$ be the probability the visitor follows a link to node w given the fact he is currently in node v. These probabilities have to be estimated from the activity measure. So basically we are constructing a flow network, where each link has an unbounded capacity.

Besides by following links, the activity measure of a blog will originate from visitors starting in a particular node. Visitors may also stop in certain blogs. This is modeled by adding two nodes *source* and *sink* to the graph and create arcs from *source* into each blog and also links from each blog to *sink*. The resulting graph is denoted as $G' = (V', A')$. The activity measure of the new nodes still needs to be defined. Of course, $r(source)$ is the number of unique visitor to the blog graph. Obviously $r(source) = r(sink)$.

We assume the flow through a link (v, w), from node v to node w, amounts to: $P(w|v) = r(v)$. When traversing a link, we assume it is more likely to take a link to a node with a higher activity measure. In other words: $P(w|v) \geq P(z|v)$ if and only if $r(w) \geq r(z)$. Based on this assumption $P(w|v)$ is defined as follows:

$$P(w|v) = \begin{cases} d(v)\frac{r(w)}{R(v)} & \text{if } v \to w \in A \\ 1 - d(v) & \text{if } w = sink \\ 0 & \text{if } w = source\,. \end{cases} \tag{4}$$

where

$$R(v) = \sum_{w \neq sink \in V' : v \to w} r(w)\,. \tag{5}$$

and $d(v)$ is a damping factor that determines the likelihood a visitor stops in a particular blog. It should hold that the sum of all probabilities equals to one, this is shown in the following proof:

$$\begin{aligned} \sum_{w \in V' : v \to w} P(w|v) &= \sum_{w \in V : v \to w} P(w|v) + P(sink|v) + P(source|v) \\ &= \sum_{w \in V : v \to w} d(v)\frac{r(w)}{R(v)} + (1 - d(v)) \\ &= \frac{d(v)}{R(v)} \sum_{w \neq sink \in V' : v \to w} r(w) + (1 - d(v)) \end{aligned}$$

$$= d(v) + 1 - d(v)$$
$$= 1$$

The flow conservation law states that incoming flow and outgoing flow, of a node v, should be equal. This should also hold for this model:

$$\sum_{w \neq sink \in V' : w \rightarrow v} P(v|w)r(w) = r(v) = \sum_{w \neq source \in V' : v \rightarrow w} P(w|v)r(v)\,. \quad (6)$$

If we look at the incoming flow we can derive some properties by looking at the following cases:
1) if $v \in V$, we conclude:

$$\frac{1}{d(v)} = \sum_{w \neq sink \in V' : w \rightarrow v} \frac{r(w)}{R(w)}\,. \quad (7)$$

2) if $v = sink$, we conclude:

$$r(sink) = r(source) = \sum_{w \in V : w \rightarrow sink} (1 - d(w))r(w)\,. \quad (8)$$

3) obviously, if $v = source$ the sum results in 0.

3.3 The algorithm

Looking back at what we have discussed so far, we have made the following choices in the context of the proposed experiment.

1. the static property, $S(p)$, will be the initial probability distribution based on ego betweenness centrality: $I = 1/ \parallel C_b \parallel_1 \times C_b$.
2. the dynamic property, $D(p)$, will be the one-step transition matrix based on the reactions (time stamps) found on the blog p.
3. The relation is defined by the Markov Model $< S, Q, L >$ where S is the set of blogs (nodes), Q is the one-step transition matrix and L is the initial probability distribution.

Based on these choices we have developed an algorithm to compute rankings on our data set. The algorithm can be divided in several steps, the first three steps are the initialization steps and the fourth step is the actual ranking computation. This is a very basic description of the steps needed in the algorithm, no optimizations have been applied.

1. Construct the one-step probability matrix from the weighted graph.
2. Compute the ego betweenness for all nodes, based on the static structure.
3. Compute the initial probability distribution from the ego betweenness values.
4. Compute the rankings for all nodes based on a history of m steps.

4 Initial Results

In this section we will very briefly cover the results we have seen so far. Based on the ideas presented in this paper we are developing a prototype. The prototype is for the most part implemented as described in this article. The current prototype can use a betweenness and (in)degree centrality measure. If we look at the indegree based initial probability distribution we see almost similar results to the ranking created by the supplier[1] of the data set. Since they also use an indegree based ranking, this should be true.

If we use the betweenness bases approach with a constant value for d, the results look promising but they have also brought a problem with the dataset to our attention. We haven't been supplied with activity data for all blogs. The data is available, therefore we expect the result to improve even further if we run the algorithm on the correct dataset and by running the algorithm with a proper implementation of the value for d.

5 Conclusion and Future Work

In order to answer the research questions, a model has been presented and it has been discussed in depth in the context of an experiment. The importance of a web site, the static component, can be measured by using any of the known centrality measures. Based on their conceptual meaning, we have chosen to primarily use betweenness centrality. In order to optimize the algorithm we have implemented an ego betweenness algorithm. The dynamic property has been defined as the number of reactions to the postings of a blog. And an approach to convert node activity into link activity has been proposed. This approach is also applicable to websites as a whole. In the most ideal situation however, we should have access to traversal information between web sites. The relation between the static and dynamic property is defined by a Markov Model, inspired by research conducted into the field of adaptive web sites. The dynamic property is used to construct the one-step probability transition matrix, Q, the static property is used to compute the initial probability distribution, L, and the states, S of the Markov Model are the unique blogs, the nodes in the graph. In order to optimize interest to a given web site, the Markov Model is used to compute a ranking based on a depth of m navigational steps. By creating links, advertising for example, to the highest ranking web sites, we can optimize interest for the given web site.

Obtaining this dynamic information is a problem. Traffic data is not freely available. We propose a solution for this problem in the domain of blogs (and possibly other community based areas). Instead of traffic we will measure reactions to a posting. This has two disadvantages. (1) The results might be polluted with 'wrong' reactions. This can be solved by improving the crawling algorithm. (2) The other disadvantage is the fact we actually need transition or traversal numbers. If we measure reactions, it's a activity measure of a node in the graph, not an arc. In order to translate the node activity numbers to traversal numbers,

[1] SiteData B.V. (www.sitedata.nl)

we made the assumption it is more likely for people to leave for a page with more visitors. Based on this assumption an approach has been presented to compute transition weights.

5.1 Future Work

Based on the foundations presented so far, some topics are still open and others raised more questions.

1. Perform the described experiment.
2. Incorporate other centrality measures for the $S(\mathrm{p})$, especially PR.
3. Extend the measuring of activity, both for blogs and for websites.
4. Research solutions to get actual traffic and/or traversal information.

References

1. Freeman, L.C.: Centrality in social networks - conceptual clarification. Social Networks **1**(3) (1979) 215–239
2. Bonacich, P.B.: Factoring and weighing approaches to status scores and clique identification. Journal of Mathematical Sociology (2) (1972) 113–120
3. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web (1999)
4. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment (1999)
5. Perkowitz, M., Etzioni, O.: Adaptive sites: Automatically learning from user access patterns. Technical report, Department of Computer Science and Engineering, University of Washington, Seattle (1997)
6. Garofalakis, J., Kappos, P., Mourloukos, D.: Web site optimization using page popularity. Technical report, University of Patras, Greece (1999)
7. Zhou, B., Chen, J., Shi, J., Zhang, H., Wu, Q.: Website link structure evaluation and improvement based on user visiting patterns. In: in HYPERTEXT '01: Proceedings of the twelfth ACM conference on Hypertext and Hypermedia., ACM Press (2001) 241–244
8. Sarukkai, R.R.: Link prediction and path analysis using markov chains. In: Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking, Amsterdam, The Netherlands, North-Holland Publishing Co. (2000) 377–386
9. Zhu, J., Hong, J., Hughes, J.G.: Using markov models for web site link prediction. Technical report, School of Information and Software Engineering, University of Ulster at Jordanstown (2002)
10. Zhu, J., Hong, J., Hughes, J.: Using markov chains for link prediction in adaptive web sites. In: In Proc. of ACM SIGWEB Hypertext, Springer (2002) 60–73
11. Eirinaki, M., Vazirgiannis, M., Kapogiannis, D.: Web path recommendations based on page ranking and markov models. In: in WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management., ACM Press (2005) 2–9
12. Everett, M., Borgatti, S.P.: Ego network betweenness. Social Networks **27**(1) (January 2005) 31–38
13. Freeman, L.C.: Centered graphs and the structure of ego networks. Mathematical Social Sciences **3**(3) (October 1982) 291–304