Utilizing scale-free networks to support the search for scientific publications

Claudia Hauff
Human Media Interaction (HMI)
University of Twente, Enschede, the Netherlands
c.hauff@ewi.utwente.nl

Andreas Nürnberger
Inst. for Knowledge and Language Processing
University of Magdeburg, Magdeburg, Germany
nuernb@iws.cs.uni-magdeburg.de

ABSTRACT
When searching for scientific publications, users today often rely on search engines such as Yahoo.com. Whereas searching for publications whose titles are known is considered to be an easy task, users who are looking for important publications in research fields they are unfamiliar with face greater difficulties since few or no indications of a publication's importance to the respective fields are given. In this paper we investigate the application of the theory of scale-free networks to derive importance indicators for a collection of publications. A tool was developed to support the user in his publication search by visualizing the publications' importance indicators derived from the number of citations received and the publication's age as well as visualizing part of the citation network structure. A preliminary user study indicates the utility of our approach and warrants further research in that direction.

Categories and Subject Descriptors
H.5.0 [**Information Interfaces And Presentation**]: General

Keywords
Information Retrieval, Scale-Free Networks

1. INTRODUCTION
When searching for scientific publications, users today often rely on search engines such as Yahoo.com. Whereas searching for publications whose titles are known is considered to be an easy task, users who are looking for important publications, e.g. publications that are fundamental or had a great influence in the research community, in research fields they are unfamiliar with face greater difficulties. Although the results of such searches are likely to contain publications in the correct research fields, few or no indications are given to the user of how important or influential the publications are to the respective fields.

Google Scholar offers such an indicator by providing the total number of received citations for each publication. This simple measurement however has a drawback: it heavily disadvantages recent publications that have not had the time yet to acquire a large number of citations.

In this paper, we present an approach, that aims to provide the user with a more valid importance indicator of publications which takes the publications' age into consideration in a principled way. It relies on the theory of scale-free networks [4] which started to emerge in the late 1990s when it became clear that many real-world networks, including citation networks, have a common property: the distribution of the number of links k connected to a node, the so-called *degree distribution* $P(k)$ of a network, follows a power-law form. This property can be described as follows: the probability of a node to have received few links from other nodes is high, while the probability of a node to have been linked to by a large number of nodes is very low. In the specific case of citation networks publications form the nodes and citations or references represent the directed links (from the citing to the cited publication) of the network. Within the last few years a number of network models were developed [2, 4, 7, 14, 19, 29] that are able to generate networks with the desired degree distribution and as a by-product closely resemble the growth process as it occurs in many real-world networks.

The knowledge gained about the true structure of real-world networks in recent years has so far been rarely exploited. Much research has concentrated on developing network models that resemble real networks as closely as possible, but few applications actually take advantage of this additional knowledge. One notable exception is the research in eradicating epidemics where the knowledge has been applied to identify highly connected nodes that should be treated first in order to decrease a virus' spreading rate [11]. We adopt a different approach and hypothesize that it is possible to gain valuable information by comparing a real-world network with its corresponding network model. The model is created from statistics derived from the real-world network such as the age of the nodes, the network size, the average degree and the degree exponent. While the degree distribution of the model and the real-world network will be the same or at least be very similar, on the individual node level the degrees will almost certainly be different. Previously, we applied this idea to the ad-hoc retrieval task of a collection of Web pages [18].

In brief, our approach works as follows: given the age of a publication and the degree distribution of the citation network under consideration, we are able to predict the expected number of citations pointing to the publication by utilizing a scale-free network model. This number is then compared with the actual number of citations the publication has received. This comparison yields an indicator of how important a publication is - if the actual number of publications citing it is higher than expected, the publication is more important than one with fewer citations than expected. This makes it possible for example, that a paper, published 12 months ago, with 5 citations pointing to it receives a higher importance score than a paper with 10 citations that was published 7 years ago.

In order to evaluate our idea, a software tool called *Visual Paper Finder* (ViPF) was developed. It allows the user to search for scientific publications and visualizes the derived importance for each returned publication as well as a part of the citation network. The user is able to navigate within the citation network, further enhancing the search process.

In a preliminary user study, the usefulness of the introduced approach and of ViPF were evaluated utilizing the collection of publications indexed by Citebase, a web service with more than 360000 publications in the fields of physics, mathematics, biology and computer science. Although the results of the user study were mixed, the general outlook was positive and warrants further research in that direction.

The remainder of this paper is organized as follows: in Section 2 the theory of scale-free networks is introduced. Section 3 presents arguments in favor of utilizing citations as importance indicators and discusses the scale-free character of citation networks. Section 4 describes ViPF in greater detail. In Section 5 the Citebase data set and the conducted user study are presented. The results of the study are described in Section 6. Finally, the conclusion and directions for future work can be found in Section 7.

2. SCALE-FREE NETWORKS

Scale-free networks appear to be abundant in natural and artificial systems and among others can be found in the social [4, 6, 23, 25], biological [21, 28] and technological [5, 16] domain. More unusual examples where one would not readily suspect a (scale-free) network structure are the network of earthquakes [1] and the medieval inquisition [24].

The most basic network model able to produce a power-law degree distribution is the Barabási-Albert (BA) model [4] which is described below. In Section 2.2 the accelerated growth model [12] is presented which is the model chosen for our experiments. It is a particular example of an extension to the original BA model. Other developments include the modeling of clustering within networks [19], of aging and physical limitations of nodes [3] and the introduction of weighted [29] or rewired links [2].

2.1 BA Model
In scale-free networks, the probability $P(k)$ of a node having k links follows a power law with degree exponent γ

$$P(k) \propto k^{-\gamma}.$$

Barabási and his collaborators identified two necessary conditions for the creation of networks with such a degree distribution: *growth* and *preferential attachment*.

The building process of scale-free networks is iterative: starting with a small number m_0 of nodes, at each time step one node with m ($m \leq m_0$) undirected links attached to it joins the network. The free ends of the new links are distributed preferentially among the nodes already in the network. Each node is denoted by its time of birth, thus node s entered the network at time s. Formally, the probability $\prod$ that node s with degree $k(s,t)$ receives a new edge at time t is defined as

$$\prod = \frac{k(s,t)}{\sum_u k(u,t)}. \qquad (1)$$

The denominator $\sum_u k(u,t)$ corresponds to the total degree of the network. Thus, the higher the degree $k(s,t)$, the higher the probability of receiving further links. Equation 1 allows us to derive a function that determines the expected number of links a node should have acquired at any time t ($t \geq s$), given the node's age s

$$k(s,t) = m \left(\frac{s}{t}\right)^{-\frac{1}{2}}.$$

Due to its simplicity, the BA model lacks many of the actions possible in real networks: neither can links be rewired or introduced between old nodes, nor can links or nodes be deleted from the network. Furthermore, the algorithm produces only undirected networks. But citation networks are directed and - as will be seen in the Experiment section - the number of links added to the citation network is not constant but accelerates as the network grows. For this reason, the model introduced next is the one chosen for the experiments.

2.2 Accelerated Growth
In directed networks, the in- and out-degree are considered separately. We will concern ourselves only with the in-degree k_{in} of a node as the number of citations received is of importance to us, not the number of citations a publication contains. In those networks, the target ends of the links are of relevance, while the source ends, which can be anywhere within or outside the network, are ignored.

A network exhibits accelerated growth when its number of links grows faster than its number of nodes, leading to a non-stationary average degree. Although negative acceleration - the number of edges grows slower than the number of nodes - is also possible, it will not be considered here.

There are two general processes that lead to accelerated growth. In the first place when the network grows the number of links a new node enters the network with can also grow. This can be the case in citation networks for example, where the amount of literature increases over time, there is more to cite and hence, the average number of references on a publication increases. A second possibility is the addition of new links between old nodes. Actor and collaboration networks can be named here.

It is assumed that the number of links grows faster than the

number of nodes according to a power-law

$$k_{in} = c_0 t^b \tag{2}$$

with b as the growth exponent and c_0 as a constant. It is clear that $b < 1$ for most real-world networks, otherwise the average degree would increase indefinitely.

If the condition $\gamma_{in} > 2$ holds for the in-link power-law distribution (as is the case for the Citebase data set), links are attached to a node with a probability proportional to

$$k_{in}(s,t) + Bc_0 t^b/(1+b)$$

with B is positive constant. This leads to

$$k_{in}(s,t) = \left(\frac{Bc_0 s^b}{1 - Bb} \right) \left(\frac{s}{t} \right)^{-(1+b)/(1+B)} - \frac{Bc_0 t^b}{1 - Bb} \tag{3}$$

for the expected number of in-links of node s with age s at time t. To summarize, in order to calculate $k_{in}(s,t)$ the following network statistics are necessary: the exponent γ_{in} of the power-law degree distribution, the accelerated growth parameters c_0 and b, the age s of each node and the total number of nodes t in the network.

For a thorough coverage of the mathematical aspects of the theory of scale-free networks and the derivation of the presented formulas as well as an in-depth look at real-world examples, the interested reader is referred to [13].

3. CITATION NETWORKS
3.1 Citations as Importance Indicators
One of the assumptions of this work is the existence of a positive correlation between the number of citations a publication receives and the publication's importance. Intuitively we expect a citation to mean that the two papers are related by content or semantics; that the cited paper is qualitatively good enough to be cited, that an author cites all papers that he should cite and none else. It is not difficult to imagine, that not all these assumptions hold in the real world - it is unrealistic to assume that a researcher knows all papers relevant to his research or that he will cite all papers that ought to be cited as there are constraints on the length of papers. There can also be other reasons for citations: a social relationship between the authors, self-citations purely to increase the citation count, negative citations (citing a paper to criticize it) or the copying of citations from other papers.

A number of studies have been conducted to determine how the citation situation in the real world differs from our expectations. If the number of citations is indeed dependent on the quality, importance or influence of a publication, one possibility to determine the validity of the assumption is to compare the citation count of high quality papers with the ones of average papers as done by Brooks [9]. He defined high-quality papers to be those that received high ratings in the peer review process. The result was that the citation count for best paper award publications was considerably higher than that of other papers. This finding implies a correlation between quality and number of citations, although it should not be forgotten that the best paper award provides a paper with a special awareness in the research community. A similar study was conducted by Rinia et al.[26], who

compared the citation count of research programs in physics in the Netherlands with peer review judgments. They also found a good - though not perfect - correlation between citation counts and peer review judgments. White et al.[27] approached the central question of this section differently. They observed a group of scientists who over a period of 10 years built up personal and professional relationships and found that professional relationships far outweigh the social ties within the group. Therefore, despite the fact that many factors influence the citation behavior of authors, overall the qualitatively good publications are likely to be cited more often than an average paper.

However, for very specific fields of research that are studied only by a small group of researchers this generalized approach most likely fails. Here, further information, e.g. about topic specificity, network cohesion or cliques have to be considered, which is, however, beyond the scope of this work.

3.2 Are Citation Networks Scale-Free?
Several researchers have investigated scientific citation networks as part of the research in scale-free networks [8, 23, 25]. With very few exceptions, only the in-degree distributions were examined and only those citations between papers that both appear in the data set were taken into consideration.

Redner [25] conducted the first large study on citation networks using publications indexed by ISI and a second data set of Physical Review D papers. The in-degree exponent γ_{in} was found to approach 3 for $k_{in} > 500$. In the regions of low k_{in} the degree distribution was following a stretched exponential. A very similar result was achieved when examining the Physical Review D data set Volumes 11-50 from the years 1975 to 1994.

A cleansed version of the SLAC SPIRES database was studied by Lehman et al.[23]. 281717 publications were included in the estimation of the degree distribution. They reported a scale-free behavior with two regimes; papers with 50 or less citations follow $P(k_{in}) \propto k_{in}^{-1.3}$ and papers with more than 50 citations $P(k_{in}) \propto k_{in}^{-2.3}$.

Boerner et al.[8] reported a best fit for their citation data from the Proceedings of the National Academy of Sciences (PNAS) not for a pure power law, but for a power law with an exponential cut-off. They suggest that the cut-off is due to the aging of papers, as the most cited papers exist less often than predicted by the pure power law form and lowly cited papers exist more often.

4. VISUAL PAPER FINDER
Imagine being faced with the task of becoming familiar with the current developments in a research area you know very little about. Apart from locating the milestone publications within the research area it is also necessary to find recent publications that have attracted much attention. Those publications are likely to contain the current state-of-the-art of the area. ViPF was developed to support users in such a scenario and is mainly aimed at researchers, PhD students, advanced undergraduate students and generally people that are new to a research field.
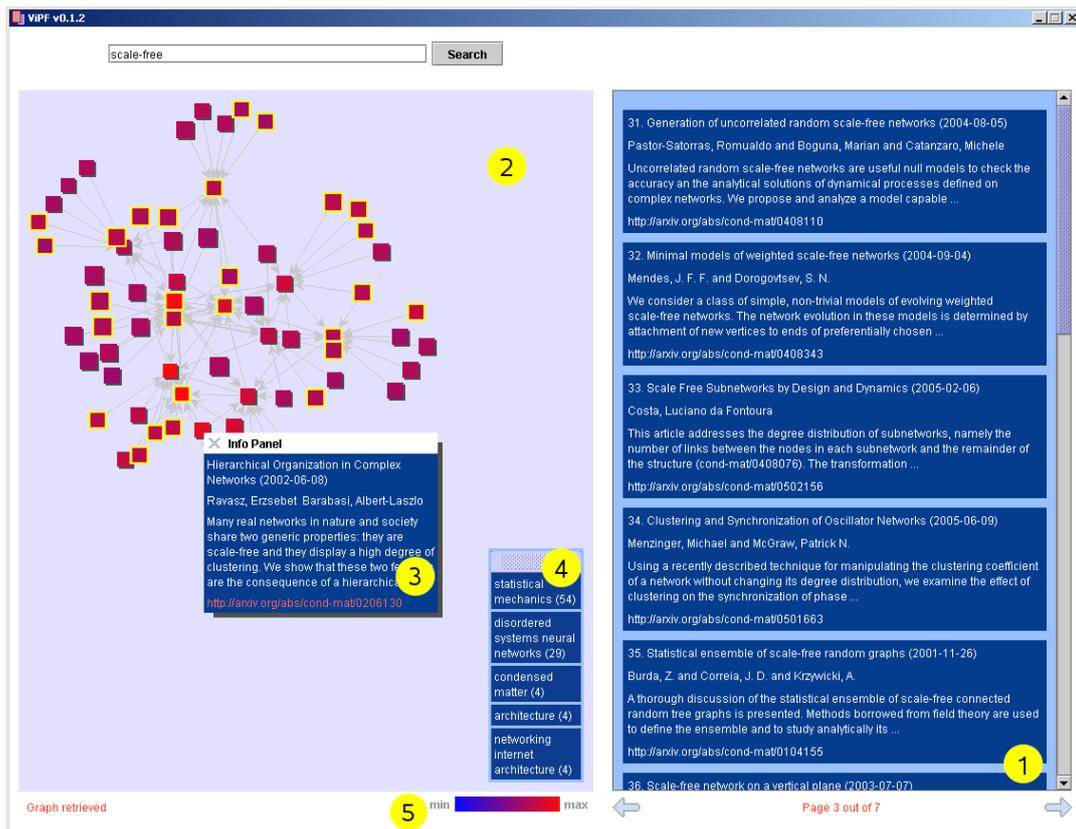
Figure 1: ViPF interface after a query was submitted: result panel (1), graph panel (2), info panel (3), subject panel (4), fitness color bar (5)

Images that visualize citation networks or more generally bibliographic networks are not difficult to find, one source is the InfoVis contest 2004 [17]. But there very few freely available tools that visualize parts of a bibliographic network interactively. CiteSpace [10] tracks the changes of a knowledge domain over time by highlighting major changes between adjacent time slices. The Growing Polygons causality visualization technique is applied in CiteWiz [15] and a multitude of different information panels with information about citations, topics and authors are presented by Paper-Lens [22]. However these tools require an intensive effort by the user since the visualizations are very complex and not feasible for everyday usage when searching for scientific publications. ViPF was developed with these problems in mind which is reflected in its simple interface.

4.1 Interface

A screenshot of ViPF's interface can be seen in Figure 1. It consists of two panels - a graph panel and a result panel. Given a query, in the result panel the ranked list of returned publications of a content-only search are presented. Each result entry consists of four parts: title and publication or first uploading date of the publication, author(s), the first part of the abstract and the URL that points to the web page where the publication can be downloaded. The graph panel visualizes part of the citation graph with a given publication as root node. After the results for a query are retrieved the citation subgraph for the top returned result is automatically shown. Clicking on an arbitrary result field retrieves the citation graph with the corresponding publication as root node. As we are interested in the publications citing a paper, the subgraph is built up by following the root node's incoming citations and doing this for every other node recursively up to a certain depth. The color of each node indicates its importance or fitness as determined by the comparison between actual and expected number of received citations. For better orientation the gradient color bar at the bottom of the graph panel shows the colors for maximum and minimum fitness. A click on any of the visualized nodes opens an info panel with information about the publication. The size of the visualized nodes varies, depending on the publication's age - the larger the node, the more recent is the publication. This feature shall make it easier to find recent papers without having to open each node's info panel to find out its publication date. The subject panel is a further help to the user. It shows the top five subjects within the retrieved subgraph. The number of nodes belonging to the respective subject is given in brackets. Clicking on a subject highlights the nodes belonging to the subject, visualized by a specially colored border. As one node can belong to more than one subject (or none), the sum of the elements in the top five subjects may exceed or fall below the total number of shown nodes.

The behavior of ViPF is managed through a parameter file. In it the age and subject indicators can be switched on or

off, the depth up to which the graph shall be displayed can be changed and the colors of all elements of the interface can be adjusted.

To avoid an overloaded display, papers with more than a certain number of citations pointing to it have not all citing papers shown. Instead, the papers were ranked according to their importance indicators and only the top n nodes were displayed. The value of n can also be modified through the parameter file.

4.2 Implementation

ViPF was implemented in Java, the graph visualization was realized with the open source libraries JGraph and JGraphAddons. ViPF relies on Citebase's retrieval engine (Xapian) and does not perform the retrieval process itself. Returned to ViPF is an XML file that contains a maximum of 100 retrieved publications. From the XML stream the required information is extracted and presented to the user in the result panel. To keep the traffic on Citebase low, the structure of the visualized part of the citation network is retrieved from a local database. The database was constructed from Citebase's metadata and includes all necessary information for the importance indicator calculation on each publication, that includes the time stamp of each paper, the number of each paper's outgoing and incoming references and a list of subjects the paper belongs to. The necessary network statistics are also stored in the database and retrieved by ViPF after every start of the program.

A question raised during the design of ViPF was whether or not to mix content scores with citation based scores in the result ranking. It was decided to use pure content ranking in order not to bias the ranking against or in favor of highly cited papers. The reasoning is, that most papers will cite the important or ground-breaking papers of a field and thus it should still be possible to gain valuable information from the visualization. Furthermore, in the graph panel it is only possible to explore papers that have received citations or reference a paper within Citebase. Since a sizable portion of papers in Citebase have no citations associated with them (as will be seen in the next section), they would then probably neither appear in the result nor in the graph panel.

5. EXPERIMENTAL SETUP
5.1 The Collection

Citebase, developed at the University of Southampton, is a search service for freely available publications of the Web. We received a citation file from the 2nd June 2005 with all available citations between publications that both appear in Citebase's index. The total number of papers amounts to 363207, with the largest proportion of papers (89%) coming from arXiv.org, an e-print repository for papers in the fields of physics, computer science, mathematics, quantitative biology and non-linear science.

The citation file contained 2501180 citations between 272349 papers, thus 25.02% of the publications have neither incoming nor outgoing references. This can be explained with the fact that Citebase only accounts for references between papers that both appear in Citebase, hence missing a substantial number of references. Although this is not an ideal
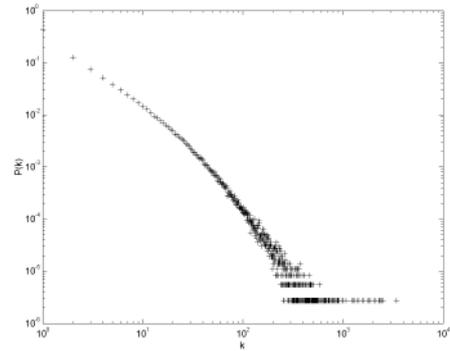
Figure 2: In-degree distribution of the Citebase data set with $\gamma_{in} = 2.1564$

situation for our experiments, at that time it was the best data set available.

5.1.1 The In-Degree Distribution

Of the papers with references, 203813 papers had received citations from other Citebase papers. Apart from the incomplete coverage of the citation network, reasons for the lack of incoming references for a paper include the recency of the paper (no author had yet the chance to reference the publication), the publication of a paper in a highly specialized field or simply the low quality of the paper.

The average number of citations is 6.89, the median is 1. 84.98% of papers have received 10 or fewer citations; only 2.55% of papers have 50 or more references pointing towards them. The top 4.21% of papers generate 50% of all incoming citations. 0.89% of citations are generated by the lowest 50% papers. These uneven numbers suggest, that the degree distribution follows a power law: there are very few highly connected nodes and many lowly connected nodes. In order to determine whether or not the Citebase in-degree distribution is indeed scale-free, $P(k_{in})$ was plotted on a log-log plot. The power law degree distribution is only defined for $k_{in} \geq 1$, the Citebase data set however contains a large portion of papers with $k_{in} = 0$. Ignoring such a large part of the collection is not an option and for this reason, the in-degree of each node was increased by 1. The resulting plot is shown in Figure 2. A slight curvature in the data set is visible. This is not unexpected though, as models are a simplification and idealization of real-world processes. The power-law form is a reasonable estimate of the observed data. We calculated the degree exponent applying the Maximum Likelihood method: $\gamma_{in} = 2.1564$.

5.1.2 Preferential Attachment

In the presented scale-free network models it is assumed that preferential attachment exists. Whether or not this is the case for the Citebase data set was investigated by applying the approach described in [20]. Let $\prod(k_{in})$ be the rate at which nodes with in-degree k_{in} receive further connections. Due to large fluctuations in the higher regions of k_{in}, the cumulative distribution function $F(k_{in})$ yields a more robust
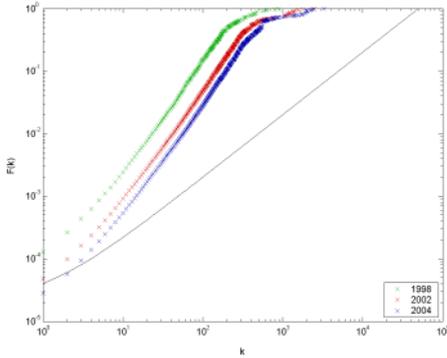
Figure 3: Preferential attachment measurements of the Citebase data set

estimate

$$F(k_{in}) = \int_0^{k_{in}} \prod (k_{in}) dk.$$

The shape of $F(k_{in})$ will indicate if preferential attachment is present. If it is absent, $\prod$ is independent of k_{in} and thus $F(k_{in}) \propto k$.

Three set of experiments in different time intervals were performed, Figure 3 shows their cumulative distributions.

The continuous line indicates the shape of a cumulative distribution that is independent of k. Clearly, the increase of $F(k_{in})$ is faster (the slope is steeper), supporting the claim that preferential attachment is at work. Moreover, the form of $F(k_{in})$ is independent of the time interval, the preferential attachment process does not change considerably over time. Although it was argued earlier that old papers should be treated differently as the age is also a determining factor for a publication's citations (older papers are less referenced), this problem is negligible for the Citebase data set as most papers in this database were written in the 1990s or later.

5.1.3 Accelerated Growth
In Table 1 the number of publications, the number of citations and the average in-degree of the Citebase data set are listed for 6 different time periods. The reason for the overall increase in degree is obvious: as more papers are made available in Citebase, the chances that references from a newly published paper point to papers already in the Citebase database increase. Recall that in the accelerated growth model it is assumed that the average degree grows as a power of t. Since the in-degree distribution was estimated by adding one in-link to each node, to keep the estimate consistent, this also happened here. At a total of 29 points in time the average degree was measured. Equation 2 was logarithmized and the parameters c_0 and b were estimated by linear regression: $c_0 = 0.1043$ and $b = 0.3439$. The knowledge of the values for γ_{in}, c_0 and b allows the calculation of the value of the only missing parameter of Equation 3: $B = 0.554$.

5.1.4 The Documents' Age
Now what remains is to assign a time stamp s to each publication of the collection. The papers were ordered by their

period	#papers	#references	$\overline{k}_{in}$
1900 - 1995	34822	86373	2.4804
1900 - 1997	71042	299754	4.2194
1900 - 1999	124447	668786	5.3741
1900 - 2001	196296	1204148	6.1343
1900 - 2003	285868	1915638	6.7011
1900 - 2005	363207	2501178	6.8864

Table 1: With an increase in network size, the average in-degree $\overline{k}_{in}$ increases.

creation or upload dates. Only 1.49% of publications had an invalid date and had to be assigned an estimated date, minimizing the impact of the erroneous or missing data. When a paper had two or more creation dates, the earliest date was chosen. The oldest paper was assigned the time stamp 1, the second oldest the time stamp 2 and so on. The youngest paper received the time stamp 363207. If two more more papers had the same creation date, their ordering was determined randomly. Papers that have neither outgoing nor incoming references were also included. The accelerated growth model does not require a node to enter the network with links attached to it.

5.1.5 The Subjects
Citebase's metadata contains one or more subject entries for 147193 (40.52%) publications. 119978 of those belong to more than one subject category. The subject entries had to be cleaned manually, since they contained entries such as 'Reviews' or 'Research Article' which were not useful for the visualization. In a number of cases one or several alphanumerical identifiers (Mathematics Subject Classification) were listed as subjects, which had to be manually converted to meaningful phrases. The final number of subjects was 4212. Using only those subjects, 140683 papers were left with one or more subjects. Those subject assignments were used in the subject panel of ViPF's interface.

5.2 User Evaluation
A preliminary user evaluation was conducted in the first two weeks of September 2005. ViPF was available for download and users were asked to test it and fill out an online questionnaire, available in German and English, afterwards.

The questionnaire was divided into three sections. The first section asked for basic information about the user, including occupation, age and field of study. The users were also asked to estimate the amount of time they had spent using ViPF. The second section consisted of 9 questions about the user's experience with ViPF. Each question had to be answered with a score between 1 (very positive/helpful/useful) and 5 (very negative/unhelpful/unuseful) or 'no opinion' respectively. The last section contained four questions about the user's view of ViPF in free-form and the users were not restricted in the length of their answers. They were asked to describe what kind of papers they had been searching for, their general impression of ViPF and what tools or web interfaces they normally use for the search for scientific publications.

| Question | #Users assigning score | | | | | | |
	1	2	3	4	5	no op.	av. score
1. How useful is the visualization of the reference graph	2	2	6	1	0	0	2.55
2. Was the tool intuitive to use?	0	7	1	3	0	0	2.6
3. How useful is the fitness indicator for each paper?	2	3	5	1	0	0	2.55
4. Inhowfar did the search results meet your expectations?	1	3	1	4	0	2	2.89
5. Do you prefer Google (Scholar) or a similar search engine over ViPF?	1	4	4	1	0	1	2.5
6. When searching within the reference graph did you consciously pick papers with a high fitness value?	4	0	3	2	1	1	2.6
7. In case the age indicator was switched on, did you find it helpful?	0	2	2	0	1	6	3
8. How familiar are you in the research area of your paper search?	1	2	3	3	1	1	3.1
9. In case the subject indicator was switched on, did you find it helpful?	1	3	1	0	0	6	2

Table 2: Evaluation results

6. RESULTS

The questionnaire was filled in by 11 users, 7 questionnaires were returned in German and 4 in English. The average age was 24.5 years, 6 users gave Germany as their home country, 2 the United States of America and one each the United Kingdom and the Netherlands. One user did not provide a country entry. The majority of users (6) were undergraduate university students, three were researchers, one a PhD student and one a high school student. All but one user who did not give any information, stated computer science or a related term as their field of study: computer science (8), information retrieval (1) and computational visualistics (1). The fact that all but one user study or conduct research in the field of computer science was reflected in the searches. Only 5 users looked for publications in fields other than computer science. The amount of time spent using ViPF was less than 1 hour for all but one tester who spent 5 hours testing it. The results for each of the 9 questions are presented in Table 2. Users that voted 'no opinion' on a question were not taken into consideration for the calculation of the average score.

The reference graph and the fitness indicator were viewed rather positively by the users (10 users gave a score between 1 and 3), although in both cases most votes (6 and 5 respectively) were given to the score 3. This undecidedness is reflected in the answer to the question how consciously the fitness indicators were used. Only 4 users answered with a score of 1, 3 users with a score of 3 (partially used) and 3 did not use the fitness indicator consciously. The age and subject indicator were accepted as a useful feature by the majority of users who responded to these 2 questions.

The replies to the question about negative aspects of ViPF were quite similar to each other. The main point of frustration was the lack of a reference graph for many searches. Due to the small number of computer science papers compared to the number of physics papers in the collection, the retrieved papers often had not received a single citation from other Citebase papers, and thus the graph panel was rendered useless. This made it difficult for the users to accurately estimate the usefulness of the fitness parameters as large reference graphs were often available in areas they had not enough knowledge in.

The visualization of the reference graph was noted as a positive aspect by 10 users in the free-form answers. The fitness indicator and the subject indicator were also positively mentioned. The simple layout of the user interface provided little distraction and was also welcomed One user summarized his thoughts about ViPF as follows: 'I guess I am so used to the ranked list kind of interface that even with conscious effort to use the other component [the graph panel], my main entry point was always the ranked list.' This factor might also have contributed to the overall results. Most users are so used to use Google (10 users listed it as search tool of their choice) that it is difficult to introduce a different system.

7. CONCLUSIONS & FUTURE WORK

One research objective of this work was to answer the question whether or not the introduced approach - deriving an importance indicator from the comparison of actual and expected number of received citations - can be utilized to support users in their search for scientific publications. This question could only be answered partially since the results of the user study were too mixed to allow a conclusion. There are several reasons for this, among them the size of the user study and the recruited users. The searches in the field of computer science often only returned a single node, so that the users had no chance to evaluate the fitness indicators properly in their field. For a thorough test, users that are experts in the various fields of physics covered by Citebase need to be recruited to give a valid estimate of the usefulness and correctness of the fitness indicator. The preliminary user study should be followed by a larger one with a clear retrieval task and a set of measurements to evaluate the users' efforts, possible with a different publication data set, to gain more representative results.

The second objective - an appropriate visualization of the citation network - was achieved. The majority of users had a positive attitude towards the presented visualization and the simplicity of the interface.

Apart from a more representative user study, there are two major directions for future work, on the one hand the im-

provement of the ViPF interface and the optimization of the retrieval and graph visualization and on the other hand the extension of the scale-free network approach.

Acknowledgments

8. REFERENCES

[1] S. Abe and N. Suzuki. Scale-free network of earthquakes. *Europhys. Lett.*, 65, 2004.

[2] R. Albert and A. Barabási. Topology of evolving networks: local events and universality. *Phys. Rev. Lett.*, 85(2):5234–5237, 2000.

[3] L. Amaral, A. Scala, M. Barthélémy, and H. Stanley. Classes of small-world networks. In *Proceedings of the National Academy of Sciences*, volume 97, pages 11149–11152, 2002.

[4] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] A. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the World-Wide Web. *Physica A*, 281:69–77, 2000.

[6] A. Barabási, H. Jeong, R. Ravasz, Z. Néda, T. Vicsek, and A. Schubert. On the topology of the scientific collaboration networks. *Physica A*, 311:590–614, 2002.

[7] A. Barrat, M. Barthélemy, and A. Vespignani. Weighted evolving networks: coupling topology and weight dynamics. *Phys. Rev. Lett.*, 92(22):228701, 2004.

[8] K. Boerner, J. Maru, and R. Goldstone. The simultaenous evolution of author and paper networks. In *Proceedings of the National Academy of Scienes*, volume 101, Supplement 1, pages 5266–5273, 2004.

[9] T. Brooks. How good are the best papers of JASIS? *Journal of the American Society for Information Science*, 51(5):485–486, 2000.

[10] C. Chen. Measuring the movement of a research paradigm. *Visualization and Data Analysis*, 5669:63–76, 2005.

[11] Z. Dezsó and A. Barabási. Halting viruses in scale-free networks. *Phys. Rev. E*, 65:055103, 2002.

[12] S. Dorogovtsev and J. Mendes. *Handbook of Graphs and Networks: From the Genome to the Internet*, chapter Accelerated growth of networks. Wiley-VCH, 2002.

[13] S. Dorogovtsev and J. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[14] S. Dorogovtsev, J. Mendes, and A. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, 85(21):4633–4636, 2000.

[15] N. Elmqvist and P. Tsigas. Citewiz: A tool for the visualization of scientific citation networks, 2004.

[16] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *SIGCOMM'99: Proceedings of the conference on applications, technologies, architectures and protocols for computer communication*, pages 251–262, 1999.

[17] J.-D. Fekete, G. Grinstein, and C. Plaisant, editors. *IEEE InfoVis 2004 Contest, the history of InfoVis*, 2004.

[18] C. Hauff and L. Azzopardi. Age dependent document priors in link structure analysis. In *ECIR*, pages 552–554, 2005.

[19] P. Holme and B. Kim. Growing scale-free networks with tunable clustering. *Phys. Rev. E*, 65:026107, 2002.

[20] H. Jeong, Z. Néda, and A. Barabási. Measuring preferential attachment in evolving networks. *Europhys. Lett.*, 61(4).

[21] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[22] B. Lee, M. Czerwinski, G. Robertson, and B. Bederson. Understanding eight years of InfoVis conferences using PaperLens. In *Posters Compendium of InfoVis 2004*, pages 53–54, 2004.

[23] S. Lehmann, B. Lautrup, and A. Jackson. Citation networks in high energy physics. *Phys. Rev. E*, 68:026113, 2003.

[24] P. Ormerod and A. Roach. The medieval inquisition: scale-free networks and the surpression of heresy. *Physica A*, 339:645–652, 2004.

[25] S. Redner. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B*, 4:131–134, 1998.

[26] E. Rinia, T. van Leeuwen, H. van Vuren, and A. van Raan. Comparative analysis of a set of bibliometric indicators and central peer review criteria. Evaluation of condensed matter physics in the Netherlands. *Research Policy*, 27, 1998.

[27] H. White, B. Wellman, and N. Nazer. Does citation reflect social structure? Longitudinal evidence from the globenet interdisciplinary research group. *Journal of the American Society for Information Science and Technology*, 55(2):111–126, 2003.

[28] S. Wuchty. Scale-free behavior in protein domain networks. *Mol. Biol. Evol.*, 18(9):1694–1702, 2001.

[29] S. Yook, H. Jeong, and A. Barabási. Weighted evolving networks. *Phys. Rev. Lett.*, 86(25):5835–5838, 2001.