# Incrementally Ranking Ephemeral Web Documents in Search Engines*

Jie Wu Karl Aberer

Swiss Federal Institute of Technology (EPF), Lausanne
School of Computer and Communication Sciences
1015 Lausanne, Switzerland
{jie.wu, karl.aberer}@epfl.ch

**Abstract**

We observe that there exists a sort of special web documents in the Internet whose life time is very short. In technical terms of search engines, these documents come into being after the web crawlers crawl the web and disappear before the next round of crawling of the search engines. We call these documents *ephemeral* web documents *in a narrow sense*. We extend this notion and call documents that newly appear (not necessarily die and disappear) in between 2 consecutive search engines' crawling efforts *ephemeral* web documents *in a broad sense*. We observe that the state-of-art search engines are hardly responsive to these special web sites and documents and are not capable of efficiently including them into search results in a timely fashion. We identify possible reasons and the problems to be solved. Finally we present some preliminary strategy to the problems. Our approach can also be used in incremental computation of PageRank.

**Keywords:** web information retrieval, search engine, ranking algorithm

## 1 Introduction

Web documents in the Internet can have many different characteristics. Some are changed or updated frequently, some seldom; some are highly demanded, some are not; some stay online for a very long time, some are pretty short-lived; some are related to everyday new or hot events, some are about topics that never evolve at all; so on and so forth. We consider these characters as semantic [1] properties of web documents which can say a lot about their features other than syntactical ones such as terms and keywords. We expect them to be helpful for search systems so that results can be selected according to the matching between these semantic document features and user interests and preferences. However, our investigation shows that even the most state-of-art search engines do little in this respect. We start with our observations on related problems and then discuss possible solutions.

### 1.1 Ephemeral Web Documents

We observe that there are a sort of special web documents in the Internet whose life time is pretty short. For example, a short-term program like a summer school sets up a web site for the convenience of the attendants and closes it immediately after the program is over. An authoring team sets up a web site for the participants to exchange ideas in order to write an application for an international academic grant and closes it after the documents are completed and submitted. Their life-span is so short that it poses a problem for modern search engines to include them into the index bases.

Most modern search engines, except the meta engines and the directory-style ones, are based on web crawlers. The typical logical trilogy of such crawler-based engines' workflow is:

1. Web crawlers crawl the web. Documents are retrieved from the Internet and stored in the server farm of the search engine.
2. Linguistic post-processing of the documents. Mainly in this step the indices of the documents are generated.

*The work presented in this paper was supported (in part) by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67322.

[1] The *semantic* used here has a meaning different from that in the notion *semantic web* which is more related to the techniques that enable automatic comprehension of the web by machines.

3. Ranking of the documents are computed. For example, Google does the costly PageRank [Page et al., 1998] computation.

This cycle takes in total a long time as the web grows larger and larger. Newly crawled and indexed documents are only available to searchers at the end of the cycle. Normally the end of one cycle also means the start of the next cycle. A typical cycle of Google is approximately one to one and a half month [Notess, 2003a] for its collection of claimed 3 billion web documents. Thus the cycle of a web site's being crawled by the crawlers is also approximately one to one and a half month. If a web site is created and linked to the Internet after the referrer has been crawled and dies before the referrer is crawled in the next round, this site is like it never existed at all in the Internet in the eyes of the web crawlers, hence the search engine!

The outcome is that even if the information on these short-lived (although one to one and a half month is not really short) web sites might be very valuable to and possibly highly demanded by other people, they don't get any chance to be exposed to and utilized by the outer web society.

In technical terms of search engines, these documents come into being after the web crawlers crawl the web and disappear before the next round of crawling of the search engines. We call these documents *ephemeral* web documents *in a narrow sense*. Ephemeral web documents may be highly demanded or may not be, for example, badly-maintained personal homepages. Similarly we have the notion of *ephemeral* web sites *in a narrow sense*.

We extend this notion and call the documents newly appear, but not necessarily die and disappear, in the middle of 2 consecutive search engines' crawling efforts *ephemeral* web documents *in a broad sense*. Similarly we have the notion of *ephemeral* web sites *in a broad sense*. For example, big unexpected events such as a war, an epidemic, a protest against some political summit conference, always give rise to such web sites and documents. They burst out in the Internet and are beyond the responsiveness of modern search engines, but they usually exist for long after their birth.

We observe the fact that the state-of-art search engines are not capable of efficiently including these special documents into their search results in a timely fashion. We then identify possible reasons and the problems to be solved, finally we present some preliminary solution to the problems.

### 1.2 Contribution of the Work

Our work addresses directly the non-linguistic semantic aspects of a search engine. With *ephemeral* documents' being integrated in the huge document and index base, the freshness of the search results is greatly improved. Users need no longer to wait for more than one month to search and find content of fresh web sites, otherwise the information may have become out-dated when it is made available on a traditional search engine.

Personalization can definitely play an important role here after the *ephemeral* documents are included. A user may choose to have results more recent, fresher or have results in a much longer time window. It's somehow like the difference between the long-term investment and the short-term investment in the stock market. Typical modern search engines take the all-in-one and one-fit-all strategy which only provide as results the documents which are highly ranked in the sense of long-term.

Other positive side effects may include the increase to some extent in respect of accuracy of results because of the enlargement of the document base that is used in the comparison of similarity between the query and the documents.

We know the cause of the problem is basically the latency of crawling cycles. Investigation shows that such typical latency is at least 4 weeks or one month [Notess, 2003a]. In addition to Google, the other examples are 4 weeks for MSN (a bit more than 1/3 size of Google) [Notess, 2003b], 1 month for AllTheWeb. Thus the question to answer becomes: How to make ephemeral documents available in a search engine as soon as possible?

As for the strategy we propose in this paper to do the rank computation, we believe it is a good example to show how the global ranking of web documents can be computed in an incremental way. Thus it is worth to put more efforts in it in the future work.

## 2 Analysis of the Problem

We particularly look into the problem that the non-stop emergence of everyday news pages has brought to the search engines. Obviously news pages can be classified as *ephemeral* web documents *in a broad sense*. Newer ones are published everyday and the general-purpose search engines are not capable of including them in time into their snapshots of the web graph in order to compute the global ranking of web documents using algorithms such as PageRank.

Simply crawling and storing them may not be the primary challenge. Given the decreasing cost of hardware and the maturity of crawler technologies, it would not be difficult for a search service provider to crawl the main news web sites around the world daily. In fact, the existence of special-purpose news search engines demonstrates the feasibility of doing so.

We claim that the main technical difficulty here is how to compute or estimate the ranking values of the *ephemeral* web documents at the global scale and the feedback to the rest of the web graph incurred by this inclusion of the set of *ephemeral* web documents.

### 2.1 A Concrete Example

We take the search for "SARS" as a concrete example in modern search engines. We know that the new epidemic SARS has drawn world wide attention since several months [Surveillance and Response, 2003]. The web sites of national and international health organizations also have drawn more and more visits ever since. We would like to see how this world trend is reflected in modern search engines.

In the Appendix we first attach the screen captures of search results from Google in early May (experiments done on 1st and 2nd of May, please refer to Appendix A).

We've thought at least more than 5 of the top 10 results would be SARS-related if not all. But to our surprise in the top 6, SARS-related web documents only are ranked no. 2 and 4 respectively. All the other results belongs to to some companies, some other organizations whose acronyms happen to be exactly or like SARS. Further looking into the cached content of the no. 2 and 4 documents, we found both of them are created on March 20, when it was 40 days ago!

We can make this observation: the PageRank used in ranking retrieved web documents can only reflect the ranking situation of 40 days ago! In other words, the searchers try to find some keyword-related documents and examine usually only the top 10 to 20 returned ones, but unfortunately the results are ranked according to the importance of them 40 days ago! This is a big problem for modern search engines, even one that is extremely successful.

But search on the same keyword on MSN Search turns out to be a quite different experience. Actually all top 15 results of MSN are about the disease SARS. This is the other surprise that we are given in this small experiment. There might be two possible explanations for this difference: firstly, MSN's collection size if only a bit more than 1/3 of that of Google, so the computation of ranking documents would be much shorter than that of Google, no matter whether MSN uses ranking algorithms similar to Google's; secondly, MSN might adjust the weights of SARS-related documents probably because statistics of queries show that "SARS" becomes a highly popular query keyword. We do not know the details here but we do think a method like this can play an important role in ranking algorithms.

Thus we can summarize the possible problems of Google as: *ephemeral* documents are not included in the collection; weights given to *ephemeral* documents are not enough or not given at all. In any case, this demonstrates the radical shortcoming of huge-scale modern search engines: it is not responsive to this sort of special *ephemeral* documents. Thus the computed ranking of the web documents is not accurate in the sense of being timely and current because the *ephemeral* documents can not play any role in the computation of global ranking of web documents.

In later sections, we will present an approach to deal with the problem in a systematic and uniform way for search engines with a huge collection of web documents like Google.

### 2.2 Comparison of Ranking Life Cycles

We observe that patterns of importance of a web document under the perspective of people and search engine ranking algorithms are different for different sorts of web documents. We call these patterns the Ranking Life Cycle of documents.

We differentiate and compare the two Ranking Life Cycles of normal and *ephemeral* documents in the following two diagrams.

For a normal document, it is unknown to a search engine, e.g. Google, until it is crawled and stored by the web crawler. Then its PageRank jumps to some non-zero value after the execution of the ranking algorithm. After that, and because of the longer and longer time of its being exposed to the web, it is linked by more and more other web pages. Some pages that point to it may disappear as well thus the number of incoming links might decrease too. However, there should be a period that the absolute number of incoming links grows. [2] Thus its PageRank grows as time flows. Then it enters a equilibrium state: no more web pages add new links to it or the decreased number almost equals to the increased number. After that, the

[2] We know the PageRank does not grow linearly along with the increasing of incoming links. To keep the explanation simple, we can make this assumption in a general sense.

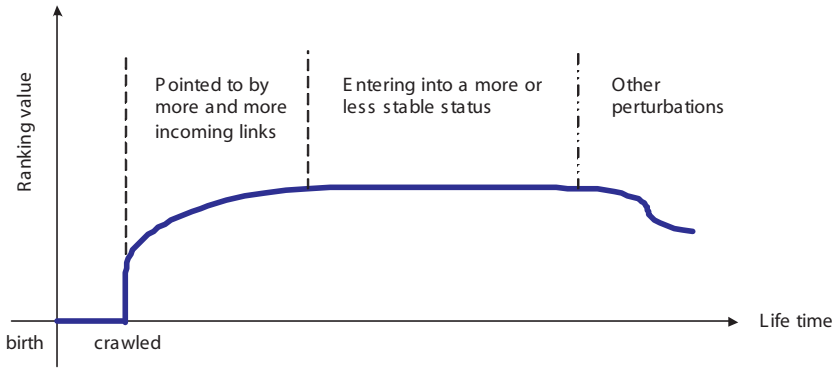


Figure 1: Ranking Life Cycle of normal documents

page becomes even older, and with the emerging of many new web documents, it is gradually forgotten by the others and thus its PageRank goes down slowly.

We would claim that for normal documents, since they are not eye-catching in general, their importance in the viewpoint of human-mind also follows such a curve.

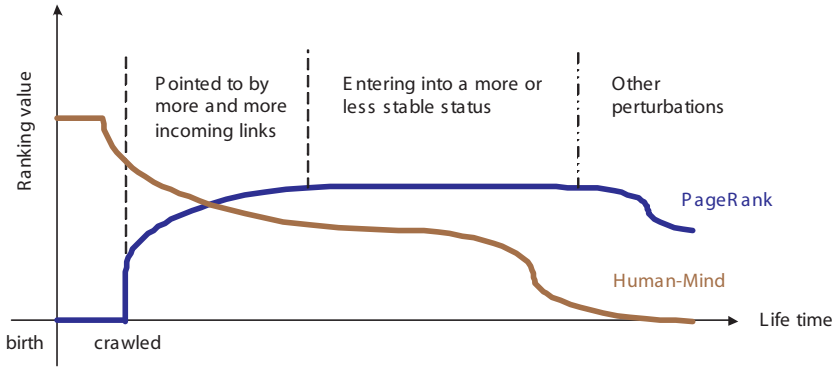


Figure 2: Ranking Life Cycle of ephemeral documents

On the other hand, the characteristics of the Ranking Life Cycle of *ephemeral* documents would be much different from above.

From the viewpoint of ranking algorithms like PageRank the curve of *ephemeral* documents would still be pretty much the same since a document is a document, and it is not any different from any other document in the eyes of the search engine. It has to be crawled again and again to be included in the ranking computation based on the matrix representation of the web graph.

For the *ephemeral* documents, the Ranking Life Cycle curve is completely different from the viewpoint of humans. Usually, such *ephemeral* documents are considered very important by and of hight interests to people immediately after they appear in the Internet. For example, web pages about an unexpected breakout of a war definitely would attract huge number of users immediately but not after the web crawlers of some search engine has crawled them, indexed them, and made them available to the end searchers, which usually happens after around 1 month.

Then as the big event develops people get accustomed to it and would not check the progress of it so frequently as they do at the beginning stage of the event. Thus the importance of such web documents drops under the viewpoint of rational humans. The longer the event lasts, the less surprise or interests would it bring. The importance diminishes and diminishes until the day comes that it is totally forgotten.

Despite these completely different characters of different sorts of web documents, they are just treated in the same way in ranking algorithms like PageRank. We think a better ranking algorithm should not simply ignore the existence of such *ephemeral* documents and only take them into consideration after they are crawled in the next round of crawling. Since they might be quite important immediately when they are born and their importance drops quickly afterwards, they should be made available to searchers as soon as possible. Actually the search results from MSN Search demonstrates the significance of doing so.

### 2.3 Delayed Reflection of Public Information Needs

Here is another observation we could derive from the experiments we made: there exists the delay between the emergence of the real world ranking of web documents and the reflection of such a shift in large-scale search engines like Google. We take the search with the same query "SARS" in late May as an example (Please refer to Appendix A). We do find that 5 of the top 10 results are about the epidemic and two of the five are no. 1 and no. 3 respectively.

While studies of Nielsen//NetRatings [Nielsen//NetRatings, 2003] show that in early April, SARS News has driven internet traffic to health-related web sites, it's not until late May that web documents about SARS finally gain sufficient recognition from the search engine. We observe there is again a delay of about 40 days or even more in reflection of the shift in public internet users' information needs.

Even though this kind of delay in a general situation is difficult to be quantified for the moment, we think a better ranking algorithm should alleviate this situation.

## 3 Three Generations of Ranking Algorithms

We categorize the ranking algorithms used by search engines so far as the following three generations. The generations are differentiated basically based on the factors that are taken into consideration when the algorithms compute the ranking of web documents.

1. Generation 1: Only or mainly textual document content are used, such as keywords/terms in the documents. Most classical models in traditional Information Retrieval and their algorithms belong to this category, such as boolean retrieval, vector retrieval, probabilistic retrieval, latent semantic indexing, etc..

   As these algorithms were studied and developed before the web came into being and had been widely used in traditional document management, the application of them to web documents directly or after some customizations does not address well the specialty of the problems of web information retrieval.

2. Generation 2: The characteristics of this generation is the use of link structure of the web. The existence of extensive hyper links between web documents marks the difference between problems in web information retrieval and the traditional ones. The link structure of the web has been proved in practice to be a very effective helper of doing the rank computation. Algorithms such as PageRank and HITS [Kleinberg, 1998] are the typical ones.

   Of course in the algorithms of the second generation, on-page factors are also used together with algorithms taking advantage of the web link structure for the search engines to get the final ranking of retrieved documents based on the query keywords.

3. Generation 3: Yet the web is not only a static graph specified by the huge number of links pointing from one document to another. It is rather a dynamic society which is comprised of both the providers, i.e. the servers which serve their web documents as the goods, and the consumers, i.e. the Internet users. The interactions between the consumers and the servers is one of the key components of the web society. Thus the preferences of the users and the viewpoints of humans towards the documents should play an important role in web-related models and algorithms.

   Thus we foresee this new generation of ranking algorithms for web IR which also takes semantic factors such as user preferences, the characteristics of documents we describe in previous sections, etc. into considerations in addition to the already used link structure and document content.

   Our work on how to integrate the *ephemeral* documents as soon as possible to a huge-scale modern search engine uniformly and seamlessly is exactly an example of finding a ranking algorithm of the third generation for the problems of web information retrieval.

## 4 Ranking Ephemeral Documents

In our preliminary work, we only look into the effect of the everyday news web documents, a set of *ephemeral* web documents *in a broad sense*. We claim that they are *ephemeral* because general-purpose huge-scale search engines have problems to integrate them into their document base in a timely fashion.

### 4.1 Existing Solutions

Existing solutions make use of special-purpose portals. Two good examples are Daypop [Daypop, 2003] and Google News [GoogleNews, 2003].

According to Daypop's about page, Daypop is a current events search engine which crawls the living web at least once a day to bring searchers the latest information relevant to users' searches. The living web the notion Daypop uses which is considered to be composed of sites that update on a daily basis: newspapers, online magazines, and weblogs which are a new form of personal journalism. Among them, newspapers give people the international headlines and weblogs give people both a subjective view of current events and a personal view of the author's life. Daypop claims that it indexes over 35,000 of the best news sites and weblogs on the net every day.

Google News claims that it presents information culled from approximately 4,500 news sources worldwide and automatically arranged to list the most relevant news first. Topics in Google News are updated continuously throughout the day, so people will see new stories each time they check the page. An automated grouping process is developed that pulls together related headlines and photos from thousands of sources worldwide which enables users to see how the same story is reported by different news organizations.

The ranking algorithms used in both of them remain commercial secrets.

### 4.2 Our Approach

We are to compute the global ranking of the enlarged web graph $W$ in a situation that global ranking of the existing web graph $G$ is already computed and newly appeared *ephemeral* web documents are crawled daily to enlarge the graph $G$. The purpose of the computation is to obtain a combined global ranking of all the documents in the enlarged graph $W$ in a incremental fashion.

We notice that reusing a converged ranking vector of a block in the start vector for PageRank is an efficient way to speed up the computation [Sepandar Kamvar and Golub, 2003a]. The authors show empirically that by making use of the block structure of the Stanford/Berkeley or a even larger web graph, a speedup by factors of at least 2 can be achieved in the PageRank computations. We believe that similar techniques can be applied in our incremental computation of the enlarged web graph after we find the proper matrix representation of the enlarged web graph.

### 4.3 Matrix Representation

We simplify the problem to solve mathematically as follows: given the web graph in a steady state, a set of newly-appeared documents is attached to it, how to compute the document ranking of the enlarged web graph?

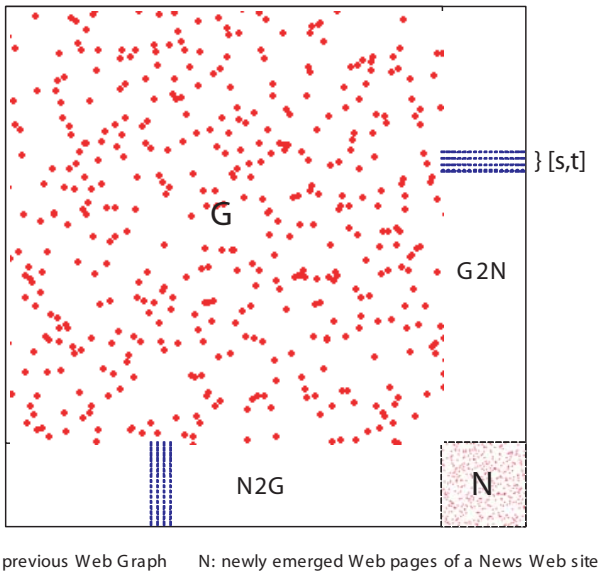


Figure 3: Matrix representation of the enlarged web graph

We show matrix representation of the enlarged web graph in Figure 3. The big rectangle tagged $G$ is the web graph before the emerging of the daily new web documents of news. Then there is the small block $N$ of news documents. The sizes of the blocks are not drawn proportional to the actual numbers of documents. The news pages would intensively link to each other because of the existence of links of relative topics, headlines, more on the same event, etc.. The existence of the several narrow rows and columns in $G2N$ and $N2G$ is because of the fact that the homepages of news web sites, the index pages of pages such as world news, sports news, etc. are presumably also crawled and included in $G$. As such pages surely have links to

their actual news entries and all those news entries normally point back to their index pages and homepages, there would be such narrow stripes in the matrix representation. As the news pages are news born there won't be other existing documents pointing to them and they are mainly self-contained and focus on their own news story so in general they do not link to other pages either except the index pages and homepages. That's why most of the space in $G2N$ and $N2G$ is empty.

We need to first review briefly the mathematical tool that is used in the computation of PageRank. A longer and more detailed version of such review can be found in a paper about adaptive methods for the computation of PageRank [Sepandar Kamvar and Golub, 2003b]. The original description of the algorithm is in [Page et al., 1998].

Taking the computation of PageRank for $G$ as an example, we let $m$ be the number of documents in the graph $G$. Let deg($u$) be the out-degree of page $u$ in $G$. Let $P$ be the stochastic transition matrix derived from $G$, where:

$$P_{ij} = \begin{cases} 0, & \text{if deg(i)= 0;} \\ \frac{1}{deg(i)}, & \text{otherwise.} \end{cases} \tag{1}$$

Let $\boldsymbol{v}$ be the $m$-dimensional column vector representing the uniform probability distribution over all documents in $G$:

$$\boldsymbol{v} = [\frac{1}{m}]_{m\times 1} \tag{2}$$

Let $\boldsymbol{e}$ be the $m$-dimensional column vector where every element $e_i$=1:

$$\boldsymbol{e} = [1]_{m\times 1} \tag{3}$$

Let $\boldsymbol{d}$ be the $m$-dimensional column vector identifying the nodes with out-degree 0:

$$d_i = \begin{cases} 1, & \text{if deg(i)= 0} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

In order to introduce the random walk that a surfer would take after visiting a page with no outgoing links, a special perturbation is added to $P$:

$$P' = P + \boldsymbol{d}\cdot \boldsymbol{v}^T \tag{5}$$

In addition the random walk model also allows that at each time step, a surfer visiting any node will jump to a random web page rather than following an outgoing link. A new matrix $P''$ is derived from $P'$ to reflect this fact where (1-$c$) is the probability of such random jumping:

$$E = \boldsymbol{e}\times \boldsymbol{v}^T \tag{6}$$

$$P'' = cP' + (1-c)E \tag{7}$$

Mathematically, the two modifications applied to the original $P$ make the new $P''$ an aperiodic and irreducible transition probability matrix of the web graph $G$. Thus the Markov chain defined by $P''$ has a unique stationary probability distribution of which the principal eigenvector of the matrix $A = (P'')^T$ is exactly the PageRank vector of $G$ to be computed.

Let $W = G \cup G2N \cup N2G \cup N$, $k$ be the number of nodes (pages) in the block of documents from news web sites, and $n = m + k$ be the total number of documents in $W$. Let $M$ be the matrix derived from $N$, where $M_{ij} = \frac{1}{deg(i)}$. There won't be a zero element here since we assume every news web pages at least points back to the homepage of the news web site. Note that $M$ alone is not a real stochastic transition matrix since elements with non-zero transition probability also exist in the columns on the left side of $M$ in $N2G$.

Suppose the index pages and homepages of news web sites are numbered sequentially from $s$ to $t$ (which can be fulfilled by applying linear transformation of the matrix $P''$), they are the only ones that point to and be pointed to by the documents in the block $N$. The number of such pages ($t - s + 1$) would be very small compared with the number $m$. Let $r_i, i \in [s,t]$ be the number of new news pages pointed by a news index page $i$. Note deg($i$)$\geq r_i$ and for the pages whose page number drop between $[s,t]$ we can safely take for granted that deg($i$)$> r_i$ because of the fact that these pages always already have outgoing links pointing to the old news pages that the web crawler has crawled.

Now we define $Q$ as (Note here deg($i$) is the new out-degree of a news index document $i$):

$$Q_{ij} = \begin{cases} P''_{ij}, & i \leq m, i \notin [s,t], j \leq m; \\ P''_{ij} - \frac{r_i}{deg(i)(deg(i)-r_i)}, & i \in [s,t], j \leq m, \exists \text{a web link } i \to j; \\ P''_{ij}, & i \in [s,t], j \leq m, \text{no such a } i \to j \text{ web link;} \\ \frac{1}{deg(i)}, & i \in [s,t], j > m; \\ \frac{1}{deg(i)}, & m < i \leq n, j \in [s,t]; \\ M_{ij}, & m < i \leq n, m < j \leq n; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

For elements in $P''$ that belong to the rows $[s,t]$ (the index pages on news web sites), we make the above special treatment by subtracting some transition probability from them and sparing to the new elements corresponding to the jumpings from these index pages to news documents. By doing so, the sum of the transition probability on the same row remains 1. We know this subtraction is possible since:

$$P''_{ij} - \frac{r_i}{deg(i)(deg(i)-r_i)} = cP_{ij} + (1-c)\frac{1}{m} - \frac{r_i}{deg(i)(deg(i)-r_i)} \quad (9)$$

$$= c\frac{1}{deg(i)-r_i} + (1-c)\frac{1}{m} - \frac{r_i}{deg(i)(deg(i)-r_i)} \quad (10)$$

$$= \frac{c\cdot m\cdot deg(i) + (1-c)\cdot deg(i)\cdot(deg(i)-r_i) - r_i\cdot m}{m\cdot deg(i)\cdot(deg(i)-r_i)} \quad (11)$$

$$= \frac{(cm+(1-c)deg(i))deg(i) - (m+(1-c)deg(i))r_i}{m\cdot deg(i)\cdot(deg(i)-r_i)} \quad (12)$$

The sufficient condition becomes:

$$r_i < \frac{cm+(1-c)deg(i)}{m+(1-c)deg(i)}deg(i) \quad (13)$$

As $m$ is the number of document in the web graph $G$, the above inequality becomes $r_i < c\,\deg(i)$ when $m \to \infty$. Thus we can choose a proper $c > \frac{r_i}{deg(i)}$ to ensure the subtraction is doable. In other words, the number of new news pages $r_i$ must be smaller than $c$ times the previous out-degree of the news index page, which can be made true easily. This conforms with the practice that $c$ is often chosen to be 0.85 in PageRank.

The nice property of $Q$ is that it is still an aperiodic, irreducible stochastic transition matrix. It is aperiodic because the matrix is still induced by the web graph. It is irreducible iff $W$ is strongly connected. This is guaranteed by the fact that, firstly, $P''$ is strongly connected; secondly, the news index pages in $P''$ are strongly connected to the news documents in $M$ since they point to each other mutually. Thus according to the Ergodic Theorem for Markov chains, the Markov chain defined by $Q$ has a unique stationary probability distribution. As the case in the classical PageRank, we can compute the global document ranking of the enlarged web graph $W$ based on the new matrix $Q$.

### 4.4 Matrix Computation

Given the matrix $B = Q^T$ and the start rank vector $\boldsymbol{w}^{(0)}$ of the documents in the enlarged web graph $W$, the standard Power Method for computing the principal eigenvector is:

function **PowerMethodRank**($B, \boldsymbol{w}^{(0)}$) {
  **repeat**
    $\boldsymbol{w}^{(l+1)} = B\boldsymbol{w}^{(l)}$;
    $\delta = \|\boldsymbol{w}^{(l+1)} - \boldsymbol{w}^{(l)}\|_1$;
  **until** $\delta < \epsilon$;
  **return** $\boldsymbol{w}^{(l+1)}$;
}

We have not studied the computation of this algorithm with a web-scale data set. But we can make some basic estimation on the cost: the size of the web in Google's collection is roughly $m = 3\ billion$. As for the average size of a news web site or a blog site, we yet to find good reference on it. A most recent relative figure we have found is from a research done in 1998 [Perrone, 1998]. It is said that in average there were about 300-500 new stories posted a day at that time. Imagining the explosion of online information, we think 400 new stories for a decent online news site is a very conservative figure for the current world. Thus the size of the matrix of the ephemeral news documents $N$ is estimated to be:

$$k = 400 \times (4,500 \sim 35,000) = 1,800,000 \sim 14,000,000 \quad (14)$$

which is $0.06\% \sim 0.47\%$ of the current size of Google document base.

As we have shown above this computation will converge to a stationary distribution because of the aperiodic and irreducible properties of the new matrix $Q$, we only consider how fast it will converge. In order for us to make use of the block structure, the computation here can take advantage of the global ranking vector $\boldsymbol{g}$ that has been computed for the matrix $G$. The start vector $\boldsymbol{w}^{(0)}$ can be set to:

$$\boldsymbol{w}_i^{(0)} = \begin{cases} \frac{m}{m+k}\cdot \boldsymbol{g}_i, & i \le m \\ \frac{1}{m+k}, & m < i \le m+k \end{cases} \quad (15)$$

Due to the huge proportional difference ($0.06\% \sim 0.47\%$) in our block structure and the presence of many 0 in $Q$ which makes it possible to apply the adaptive methods for PageRank computation [Sepandar Kamvar and Golub, 2003b], we expect this computation to converge in a very small number of rounds. To do empirical calculations on real data of the web snapshot crawled by web crawlers is our work of next step. We expect to have promising results based on the analysis we have done above.

## 5 Related Works

There is the work [Serge Abiteboul and Cobena, 2003] where the estimate of page importance is refined continuously as the web graph is visited and crawled gradually. It is very suitable for focussed crawling to the most interesting pages as the authors commented in the paper. But as the knowledge of the web graph is incomplete when the computation is done on-line, the results would not be that accurate compared with those of the off-line approaches based on the complete web graph.

Except that, currently we don't see any similar work is going on to address the problem elaborated in this paper. Several reasons may account for this situation:

1. Most crawler-based search engines continue with their trilogy of roughly monthly crawling of the whole web, linguistic post-processing, rank computation. It seems that people have not realized the fact that these special web documents require special treatment so that they can be integrated in the big umbrella of a general-purpose search engine in a uniform fashion.

2. People may not consider it really important to solve this problem. The current centralized strategy is good and enough.

3. Separate solutions and systems are provided to address the problem, for example, news.google.com, which by the way likely to mainly use keyword based algorithm to compute the similarities. This approach might be suitable for different categories of Internet documents, for example, newsgroup posts which have a completely different logical document structure; or images which is basically in binary file format and simply can't be treated like a plain text file. But why even the same web hypertext files are processed and served to searchers in separate engines? This does not make much sense although it is a good transition solution.

Thus our work tries to fit in this gap between the general-purpose search engine and a separate engine of dealing with only small part of the web document set, to find a way to integrate and rank all web documents, no matter special or not, seamlessly in a huge general-purpose search engine.

## 6 Conclusion

In this paper, we first identify the existence of a sort of special documents in the web and define the *ephemeral* documents with regard to the operations of web crawlers of search engines. We analyze the difficulty and necessity of including *ephemeral* documents as soon as possible to a search engine's candidate list of documents to serve users' queries. We then propose an strategy that makes it possible to compute the ranking of these *ephemeral* documents as part of the global web document ranking. By doing so, the ranking of normal and ephemeral documents can be unified seamlessly. However we admit that this is only the preliminary effort to solve this problem and we have yet to make use of the special structure of the newly defined matrix $Q$ of the enlarged web graph to see how to speedup the computation.

In addition, the approach provides strong support to a decentralized architecture [Wu, 2003] for web and peer-to-peer search engines as the incrementally emerged *ephemeral* documents such as pages on news web sites can be integrated in the big picture incrementally and in a step-by-step, decentralized way.

Furthermore, this approach poses no contradiction to using separate solutions. For example special search portals similar to news.google.com can be easily built upon such a unified ranking scheme out of a general-purpose huge-scale search system by only retrieving the news documents.

In the future, we think it's a good idea to extend the study on the effects of semantic factors on rankings, which would lead to promising new algorithms that belong to the third generation according to our categorization. We expect that by combining more such factors in ranking algorithms, search engines can provides fresher results that better satisfy the users' information needs.

More types of such special *ephemeral* web documents, *in a narrow sense* or *in a broad sense*, will be identified and studied to see their influences on the global ranking of web documents in a search engine. In addition to analytical work similar to what is presented in this paper, we will also carry out large-scale real world crawling of web documents to verify our ideas in practice.

## References


[Daypop, 2003] Daypop (2003). URL `http://www.daypop.com/info/about.htm`. (Date visited: May 27, 2003).

[GoogleNews, 2003] GoogleNews (2003). URL `http://news.google.com/help/about_news_search.html`. (Date visited: May 27, 2003).

[Kleinberg, 1998] Kleinberg, J. (1998). Authoritative Sources in a Hyperlinked Environment. In: *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.

[Nielsen//NetRatings, 2003] Nielsen//NetRatings (2003). URL `http://www.nielsen-netratings.com/news.jsp?section=new_pr&thetype=date&theyear=2003&themonth=3`. (Date visited: May 27, 2003).

[Notess, 2003a] Notess, G. R. (2003a). Search Engine Statistics: Freshness Showdown. URL `http://www.searchengineshowdown.com/stats/freshness.shtml`. (Data from May 17, 2003. Date visited: Jun. 30, 2003).

[Notess, 2003b] Notess, G. R. (2003b). Search Engine Statistics: Relative Size Showdown. URL `http://www.searchengineshowdown.com/stats/size.shtml`. (Data of Dec. 31, 2002. Date visited: Jun. 30, 2003).

[Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Tech. rep., Stanford University.

[Perrone, 1998] Perrone, J. L. (1998). Spotlight on the Beeb. URL `http://www.ojr.org/ojr/workplace/1017967815.php`. (Date visited: May 27, 2003).

[Sepandar Kamvar and Golub, 2003a] Sepandar Kamvar, C. D. M., Taher Haveliwala and Golub, G. (2003a). Exploiting the Block Structure of theWeb for Computing PageRank. Tech. rep., Stanford University.

[Sepandar Kamvar and Golub, 2003b] Sepandar Kamvar, T. H. and Golub, G. (2003b). Adaptive Methods for the Computation of PageRank. Tech. rep., Stanford University.

[Serge Abiteboul and Cobena, 2003] Serge Abiteboul, M. P. and Cobena, G. (2003). Adaptive On-Line Page Importance Computation. In: *Proceedings of World Wide Wed Conference 2003 (WWW2003)*. Budapest, Hungary.

[Surveillance and Response, 2003] Surveillance, C. D. and Response (2003). Severe acute respiratory syndrome (SARS): Status of the outbreak and lessons for the immediate future. Geneva, Switzerland. URL `http://www.wto.int/`. (Date visited: May 27, 2003).

[Wu, 2003] Wu, J. (2003). Towards a decentralized search architecture for the web and P2P systems. In: *Proceedings of the Workshop on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2003), the fourteenth conference on Hypertext and Hypermedia, HyperText 2003*. Nottingham, U.K.


# APPENDIX

## A Experiments with Search Engines

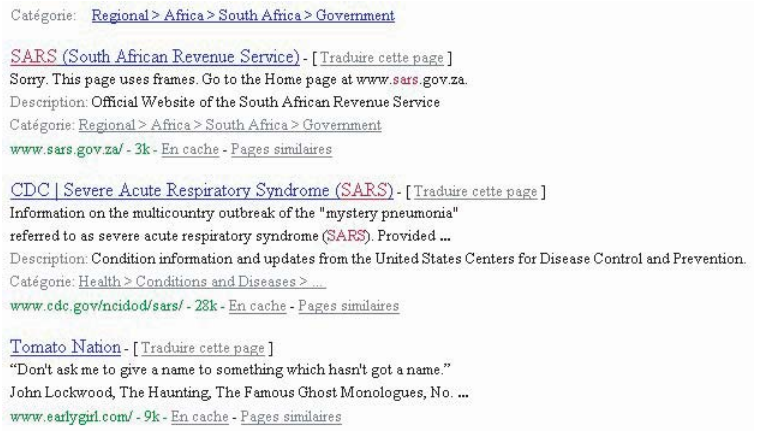


Figure 4: Top results 1-3 from Google in early May

The captured screens here are the snapshot of search results with the keyword "SARS" (Severe Acute Respiratory Syndrome) at ca. 21:45, 1.5.2003, Thu. from Google, ca. 23:05, 1.5.2003, Thu. from MSN

Search, ca. 10:15, 2.5.2003, Fri. from Google News, and finally ca. 19:07, 27.5.2003, Tue. from Google. Not all screens are shown here due to the limit of space.

In the top 3 results from Google only the second result is about the medical SARS.

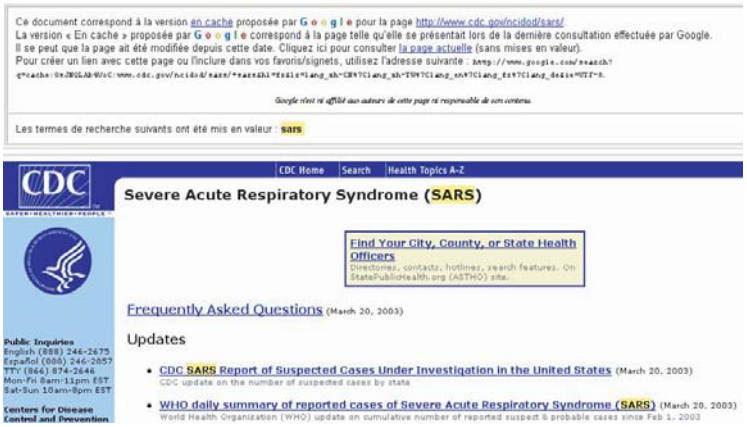


Figure 5: Cached content of the result no. 2 in early May

The cached content of the no. 2 document was actually crawled and saved on Mar. 20, 2003.