<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Feature-conscious ranking framework</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Xu, Kaijian</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Xu, K. (2007, March). Feature-conscious ranking framework. Presented at Discover URECA @ NTU poster exhibition and competition, Nanyang Technological University, Singapore.</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>2007</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10220/9058">http://hdl.handle.net/10220/9058</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>© 2007 The Author(s).</td>
</tr>
</tbody>
</table>
The PageRank algorithm [1] for Web search ranking models the Internet as a content-agnostic link graph, and derives the importance of individual Web pages without regard to their content or other properties. Various modifications of PageRank to take Web page features into account have been proposed [2], but most variants are intended for specific applications and cannot be generalized to handle any page features. Hence, we propose the Feature-Conscious Ranking framework (FeatRank) as a universal framework to incorporate any generic page feature into the page ranking computation.

Preliminaries: PageRank

Given a directed link graph with $N$ Web pages, let $A$, $E$, and $N$ be the link adjacency matrix, the link matrix, and the number of Web pages, respectively. Then, the PageRank equation is:

$$ r = (A + (1 - \gamma))E r $$

where $r$ is the column vector of PageRank scores of each page, $A$ is the link adjacency matrix, $E$ is the link matrix where $e_{ij}$ is the probability that a Web surfer will remain at the same node during a random walk of the link graph, and $\gamma$ is the probability of teleporting to and from any page.

The FeatRank Framework

The FeatRank framework, defined by $\alpha$, $\rho$, and $\gamma$, generalizes a family of PageRank variants that augments the link graph in order to explicitly account for page features in its rank computation. Any quantifiable link-independent scalar property associated with a Web page is admissible. A self-loop with link weight $\gamma$ corresponding to the page feature is added to each "featured" page. If $\gamma = 0$, the score is propagated to and from all pages.

Intuitively, $\gamma$ denotes the probability that a Web surfer will remain at the same node during a random walk on the link graph. This value must be subtracted from the probabilities to allow for teleporting to and from any page, as illustrated in Figure 1. We assign a parameter $\gamma$ to denote the ratio of $\gamma$ consumed from the self-loops; the rest is consumed from teleportation.

The FeatRank equation is thus given by:

$$ r = (A + \alpha + \gamma S + \rho I + (1 - \gamma))E r $$

where $L = \alpha + \gamma S$, where $\gamma = |E|$, $I = (1 - \gamma)E + \rho I$, $S = diag(\gamma_1, \ldots, \gamma_N)$, and $\gamma_i$ is the probability that a Web surfer will remain at the same node during a random walk of the link graph.

FeatRank Properties

Fixing $\gamma$ to the commonly used value of 0.85, we focus on the effects of varying parameters $\alpha$ and $\rho$ on featured pages and the ranking ecosystem as a whole. The effects a controlled change in the ranking of featured pages. The score $r$ of a page $i$ varies monotonically with respect to the weight of its self-loop $\gamma$:

- If $\gamma > 0$, $r$ increases monotonically with $\gamma$ (score amplification)
- If $\gamma < 0$, $r$ decreases monotonically with $\gamma$ (score attenuation)
- If $\gamma = 0$, $r$ remains unchanged.

$\alpha$ controls the propagation of the score amplification/attenuation of a featured page to nearby reachable pages in the link graph. Given a featured page with a self-loop, the propagation of its score amplification/attenuation is generally inversely proportional to $p$.

- If $p > 0$, the scores of nearby reachable pages will be amplified/attenuated.
- If $p < 0$, the scores of nearby reachable pages will be demoted.
- If $p = 0$, no propagation, relative ordering of other pages remains unchanged.

By fine-tuning the parameters, FeatRank can be adapted to a variety of different applications. As shown in Table 1, there are altogether 7 possible operating configurations for FeatRank.

Existing Ranking Models as Special Cases of FeatRank

The generic nature of FeatRank can be demonstrated by comparing it to existing ranking models such as PageRank and its variants. In particular, it can be shown that:

- PageRank is a special case of PageRank with $\gamma = 0$.
- PageRank with post-processing is a special case of PageRank with $\alpha = 0$.
- Personalized PageRank is a special case of PageRank with $p = 0$.

Application of FeatRank

The procedure of applying FeatRank to practical ranking systems is as follows:

- Identify the feature to apply, e.g., "featured article" in Wikipedia.
- Train values for $\alpha$ and $\rho$ using a parameter optimization algorithm to minimize difference between the rank list generated by FeatRank and partial ranking lists obtained from domain experts.
- Review the implications of the parameters. $\gamma$ corresponds to the "strength" the feature has for rank basing – a large absolute value of $\gamma$ indicates a useful feature, while close to zero suggests that the feature may be discarded. If $\gamma$ is close to $\omega$, the effect of the feature will be limited to the page itself. Hence, a PageRank with proper post processing may suffice.

- Apply FeatRank for ranking.

Experimental Results

We conducted experiments to (1) verify the properties of FeatRank; evaluation of the parameters' effect on ranking; (2) explore actual application of FeatRank to bias ranks of featured pages and pages surrounding them.

We used two subsets of articles from the English language Wikipedia "stable release" (26 Aug 2008). The first subset, wkiWM, consists of all 159 articles under the page "Web Browsers" and is used to verify the properties of FeatRank. The second subset, webWW, consists of 546 articles from the category "Microsoft Windows" and is used in our pilot application of FeatRank. Both datasets were manually pre-processed to remove irrelevant articles. The link graph for each subset was built from the navigational structure between the individual articles on Wikipedia.

We choose Wikipedia as our dataset because it has well-defined features for Web pages (articles), enabling us to enable FeatRank in a real environment. To provide quality control, Wikipedia defines a grading scheme for its content. We treat pages in wkiWM and wkiWM that are rated "B-Class" and above as "good articles" and make use of featured articles ("FA-Class") as our self-loop feature.

We define the following metric to evaluate the rank promotion/demotion of FeatRank over PageRank:

$$ \text{Gain} = \text{FeatRank(PageRank)} - \text{PageRank(PageRank)} $$

where $\text{Gain}$ is the rank gain of FeatRank over PageRank. To verify the properties of FeatRank, we use a self-loop to “Mozilla Firefox”, the only article in wkiWM with Wikipedia “featured article” status, and evaluate the changes in each group of pages at a defined number of hops away in the link graph, denoted by $D$. Using 3 representative $\gamma$ values and fixing $\alpha = 0.85$, we vary $\gamma$ over $(0.85,0.85)$ to study its effect on the Gain ratio. The results (Figure 2) show that when $\gamma = 0$ the featured article “Mozilla Firefox” was promoted, and when $\gamma = 0$ it is demoted; (2) the effect on the featured article is independent of $\gamma$.

Verifying the Properties of FeatRank

To verify the properties of FeatRank, we use a self-loop to “Microsoft Firefox”, the only article in wkiWM with Wikipedia “featured article” status, and evaluate the changes in each group of pages at a defined number of hops away in the link graph, denoted by $D$. Using 3 representative $\gamma$ values and fixing $\alpha = 0.85$, we vary $\gamma$ over $(0.85,0.85)$ to study its effect on the Gain ratio. The results (Figure 2) show that when $\gamma = 0$ the featured article “Mozilla Firefox” was promoted, and when $\gamma = 0$ it is demoted; (2) the effect on the featured article is independent of $\gamma$.

Exploring Real Applications for FeatRank

As a pilot study, we used FeatRank to induce a ranking of articles in the wkiWM dataset. We treat “good articles” rated "B-class" and above in Wikipedia as the ground truth, similar to ranking obtained from domain experts. Articles in the same class are assigned the same ranking. We then evaluate the efficacy of FeatRank in identifying “good articles” given the seed set of “featured articles” by comparing its performance against PageRank using the following metric:

$$ \text{Gain} = \text{Gain(FeatRank,PageRank)} $$

Next, using 2 representative values of $\gamma$, we vary $\alpha$ in $[0.05,0.95]$. Figure 3 shows that (1) the effect on surrounding pages is greatest with small $\gamma$; the effect is gradually diminished when $\gamma$ is increased; and (2) the gain on the featured article is independent of $\gamma$.

School of Computer Engineering

Project Title: FeatRank: Feature-Conscious Ranking Framework

Student: Kaijian Xu

Supervisor: Asst Prof Zhang Jun

Collaborators: Asst Prof Chang Kuiyu, Qi He