Collaborators: Asst Prof Chang Kuiyu, Qi He

Student: Kaijian Xu

Project Title: FeatRank: Feature-Conscious Ranking Framework

Feature-Conscious Ranking Framework

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 feature. 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 $(p_1, p_2, \ldots, p_N)$, the PageRank equation is:

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

where $r = \{r_1, r_2, \ldots, r_N\}$ is the PageRank score vector, $A$ is the link adjacency matrix where

$$ a_{ij} = \begin{cases} 1 & \text{if } p_i \text{ links to } p_j \\ 0 & \text{otherwise} \end{cases} $$

where $\alpha \in [0,1]$ denotes probability of traversing an outline, $E = \{e_{ij}\}$ is the permutation matrix where $e_{ij} = \text{probability of teleportation from } p_j \text{ to } p_i$.

Usually $E = \gamma I$ where $I$ is a column vector comprising all ones, giving a uniform probability of teleporting to and from any page.

The FeatRank Framework

Our work, FeatRank framework, generalizes a family of PageRank variants that augments the link graph in order to explicitly account for page features in its ranking computation. Any quantifiable link-independent scalar property associated with a Web page is admissible. A self-loop with link weight $\gamma$ corresponding to the feature is added to each "featured" page.

Intuitively, $\gamma$ denotes the probability that a Web surfer will remain at the same node during a random walk of the link graph. This value must be subtracted from the probabilities of following an outline and/or random teleportation so that all these probabilities sum up to one as illustrated in Figure 1. We assign a parameter $\gamma$ to denote the ratio of $\gamma$ consumed from the outlinks; the rest is consumed from teleportation.

The FeatRank equation is thus given by:

$$ r = (LA + \frac{1-\gamma}{N} S) r $$

where

$$ L = \gamma^{-1} (A + \frac{1}{\gamma} I), \quad \gamma \in [0,1] $$

and

$$ S = diag(\gamma_1, \gamma_2, \ldots, \gamma_N) $$

where $\gamma_i \in [0,1]$.

FeatRank Properties

Fixing $\alpha$ to the commonly used value of 0.85, we focus on the effects of varying parameters $\gamma$ and $\rho$ on featured pages and the ranking ecosystem as a whole.

Effective a controlled change in the ranking of scored features. The score $r_i$ of a page $p_i$ varies monotonically with respect to the weight of its self loop $\gamma$:

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

$\rho$ 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 $\rho$ with the following properties:

- For $\rho > 1$, the scores of nearby reachable pages will be amplified/attenuated.
- For $\rho < 1$, the scores of nearby reachable pages will be amplified/attenuated.
- For $\rho = 1$, 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 FeatRank with $\gamma = 0$.
- PageRank with post-processing is a special case of FeatRank with $\gamma = \alpha$.
- Personalized PageRank is a special case of FeatRank with $\rho = 0$.

Application of FeatRank

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

1. Identify the feature to apply, e.g. "featured article" in Wikipedia.
2. Train values for $\gamma$ 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.
3. Review the implications of the parameters. $\gamma$ corresponds to the "strength" the feature has on ranking based on a large absolute value of $\gamma$ indicates a useful feature, while close to zero suggests the feature may be discarded. If $\gamma$ is close to $\alpha$, the effect of the feature will be limited to the page itself. Hence, a PageRank with proper post processing may suffice.
4. 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, wikiWM, consists of all 159 articles under the category "Web Browsers" and is used to verify the properties of FeatRank. The second subset, wikiMW, 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 evaluate FeatRank in a real environment. To provide quality control, Wikipedia defines a grading scheme for its content. We treat pages in wikiWM and wikiMW 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:

$$ \text{Gain(FeatRank, PageRank)} = \frac{\sum_i \text{Gain}(p_i)}{N_{\text{articles}}} $$

where $\Gamma$ and $\Phi$ represent the ordinal ranking induced on the $P$ page by PageRank and FeatRank respectively. Note that we use a smaller evaluation page set of size $M$ here, instead of the whole link graph. Because the average page rank changes over the whole link graph will always converge to zero, it is more meaningful to measure only subsets of pages whose ranks are expected to change.

Verifying the Properties of FeatRank

To verify the properties of FeatRank, we use a self-loop to "Mozilla Firefox", the only article in wikiWM 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 $D(x)$. 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" is promoted, and when $\gamma = 0$ it is demoted; and (2) the effect on the featured article is independent of $\rho$.

Exploring Real Applications for FeatRank

As a pilot study, we used FeatRank to induce a ranking of articles in the WikiWM 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 rating. 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:

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

The evaluation conducted on all "good articles" with $\gamma$ in $[0.2, 0.8]$ and $\rho$ in $[0.1, 0.85]$ (Figure 5) showed that FeatRank is consistently better than PageRank. The best configuration is with small $\gamma$ and large $\gamma$. With $\text{Gain}(\text{PageRank, GroundTruth}) = -22.0847$, FeatRank yields a dramatic 28.28% reduction in rank error compared to PageRank.

References


We thank the reviewers for their helpful comments and suggestions.