Measuring Similarity to Detect Qualified Links

Xiaoguang Qi, Lan Nie, and Brian D. Davison
Dept. of Computer Science & Engineering
Lehigh University
Traditional link-based ranking algorithms (PageRank, HITS, and most variations)

Links imply merit of the target page?

Is this always true? No!
• Spam links
• Navigational links
• Advertising links
• Other irrelevant links
• **These links**
  – may be useful for humans;
  – are effectively noise for link analysis.

• **Traditional link analysis algorithms do not distinguish them from useful links.**
  – As a result, the target pages of these links could get unmerited higher ranking.
• “Qualified links”
  – Links that are qualified to make a recommendation regarding the target page

• Our proposed approach
  1. Identify and filter out “unqualified links”
  2. Perform link analysis on the reduced link graph

• In our experiments, this approach can boost ranking performance.
Background

• Hyperlink-Induced Topic Search (HITS) [Kleinberg 1997-1999]
  – The score of a hub (authority) depends on the sum of the connected authorities (hubs)

\[
A(p) = \sum_{q:q\rightarrow p} H(q) \quad \text{and} \quad H(p) = \sum_{q:p\rightarrow q} A(q)
\]

• Bharat and Henzinger (1998)
  – A number of improvements to HITS
    – \textit{imp}
      • Re-weight links involved in mutual reinforcement
      • Drop links within the same host
**PageRank**

- Random surfer model

\[
PR(i) = (1 - d) \sum_{j: j \rightarrow i} \frac{PR(j)}{O(j)} + d \frac{1}{N}
\]
Related work

- Chakrabarti et al. (1998) improving HITS by adjusting the weights of links according to their anchor text and surrounding text.
- Lempel and Moran (2000) defined the tightly-knit community (TKC) effect.
- Li et al. (2002) pointed out the small-in, large-out communities could dominate HITS results.
- Wu and Davison (2005) proposed a two-step algorithms to identify link farms.
- Benczur et al. (2006) proposed to detect nepotistic links using language models.
- Carvallo et al. (2006) proposed to detect noisy links at the site level by examining link structure among web sites.
• Introduction
• Approach
• Experiments
• Discussion & Conclusion
Approach AIRWeb 2007

- original web graph
- human label
- Classifier
- train
- a sample web graph
- human label
- human-labeled links
- labeled web graph
- classify
- Link Analysis
- filter
- reduced web graph
Welcome to The Applegate Directory

If you have come here from any domain other than www.applegate.co.uk and wish to visit the Applegate Directory then please click here.

Our SEO work has helped generate a 450% boost in traffic across all our sites in the twelve months. Find out why.
Spam links

How to determine whether a link is qualified?

• There are many features that can be used.

• In this preliminary work, we studied six similarity measures between the source and target pages.
  – Hostname, URL, topic vector, tfidf content, anchor text, non-anchor text
Similarity measures

• **Hostname similarity**
  – The portion of common substrings of two hostnames

\[
\text{Sim}_{\text{host}}(x,y) = \frac{2 \times | \text{Substr}(\text{host}_x,r) \cap \text{Substr}(\text{host}_y,r) |}{| \text{Substr}(\text{host}_x,r) | + | \text{Substr}(\text{host}_y,r) |}
\]

• **URL similarity**
  – Analogous to hostname similarity
  – The portion of common substrings of two URLs
Similarity measures (Cont.)

• **Topic vector similarity**
  - Cosine similarity of two topic vectors
    \[
    \text{Sim}_{\text{topic}}(x,y) = \sum_{i=1}^{n} v_{x,i} \times v_{y,i}
    \]
  - A topic vector \( v_x = (v_{x,1}, v_{x,2}, \ldots, v_{x,n}) \)
  - Each component is the probability that page \( x \) is on topic \( t \)
  - Can be computed using a classifier

• **Tfidf content similarity**
  - Cosine similarity of the two tfidf vectors
    \[
    \text{Sim}_{\text{content}}(x,y) = \frac{\sum_{t \in T} (x_t \times y_t)}{\sqrt{\sum_{t \in T} x_t^2} \times \sqrt{\sum_{t \in T} y_t^2}}
    \]
Similarity measures (Cont.)

• Anchor text similarity
  – Analogous to content similarity
  – Similarity (tfidf cosine) of anchor text on the two pages

• Non-anchor text similarity
  – Analogous to content similarity
  – Similarity of non-anchor text on the two pages
• Introduction
• Approach
• Experiments
• Discussion & Conclusion
Datasets

- **Query-specific datasets**
  - Collected and used by [Wu and Davison 2005]
  - 58 queries used in previous research, from ODP category names, and popular queries

- **Global dataset**
  - A 2005 crawl from the Stanford WebBase
Link classification

• Human labeling of links
  – More than a thousand links randomly selected from five query-specific datasets
  – Two human editors

• Link classification based on all features
  – A linear SVM classifier using $\text{SVM}^{\text{light}}$
  – Two fold cross validation
  – Average accuracy 83.8%
Anchor text similarity was found to be the most discriminative feature.
Qualified HITS

- Performed on query-specific datasets.
- Unqualified links, identified by the classifier, are removed.
- $imp$ is performed on the reduced link graph.
- In this experiment, 37% of the links are removed.

---

Experiments AIRWeb 2007
Qualified PageRank

- Performed on WebBase dataset.
- Unqualified links, identified by the classifier, are removed.
- *PageRank* is performed on the reduced link graph.
- In this experiment, 0.4% of the links are removed.

![Graph showing comparison between PR and Q-PR]

- Score@10 for PR: 0.64
- Score@10 for Q-PR: 0.68
- Precision@10 for PR: 0.58
- Precision@10 for Q-PR: 0.60
• Introduction
• Approach
• Experiments
• Discussion & Conclusion
Conclusion

• Summary of approach
  – Remove noise (unqualified links) from link graph before link analysis is applied
  – Identify unqualified links by measuring similarities between their source and target pages

• Qualified HITS is able to improve precision by 9% over Bharat and Henzinger’s *imp.*
Discussion

• How to determine the qualification of a link is an open question.
  – We used similarities between source and target page to demonstrate the potential of the idea.
  – What about other similarity measures?
  – What about non-similarity features?

• A closer look at link classification.
  – We trained a multi-class classifier to distinguish between qualified, spam, navigational, advertising, and other irrelevant links.
  – The classifier is effective in finding spam links
    • Not very helpful for other types.
Thank you!

WUME Laboratory
http://wume.cse.lehigh.edu/