Measuring Similarity to Detect Qualified Links

Xiaoguang Qi, Lan Nie, and Brian D. Davison

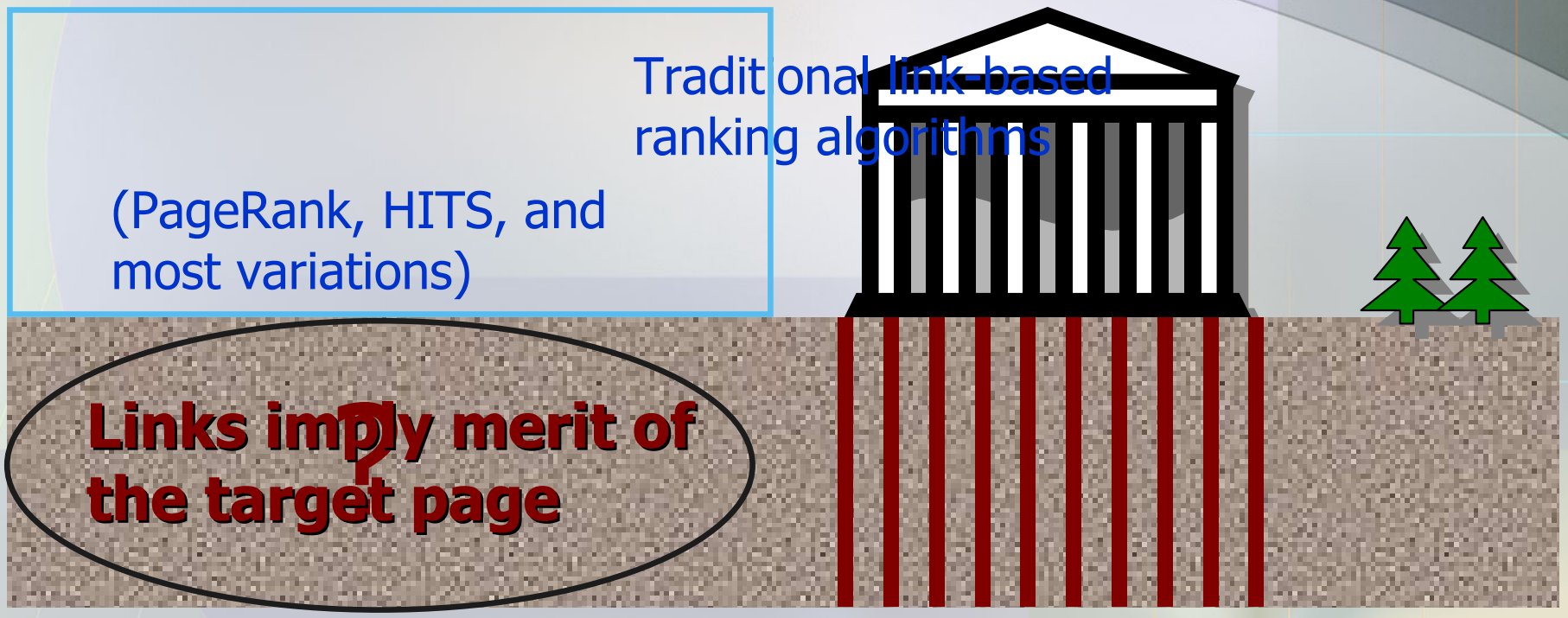
Dept. of Computer Science & Engineering
Lehigh University

LEHIGH
UNIVERSITY

- **Introduction**
- **Approach**
- **Experiments**
- **Discussion & Conclusion**

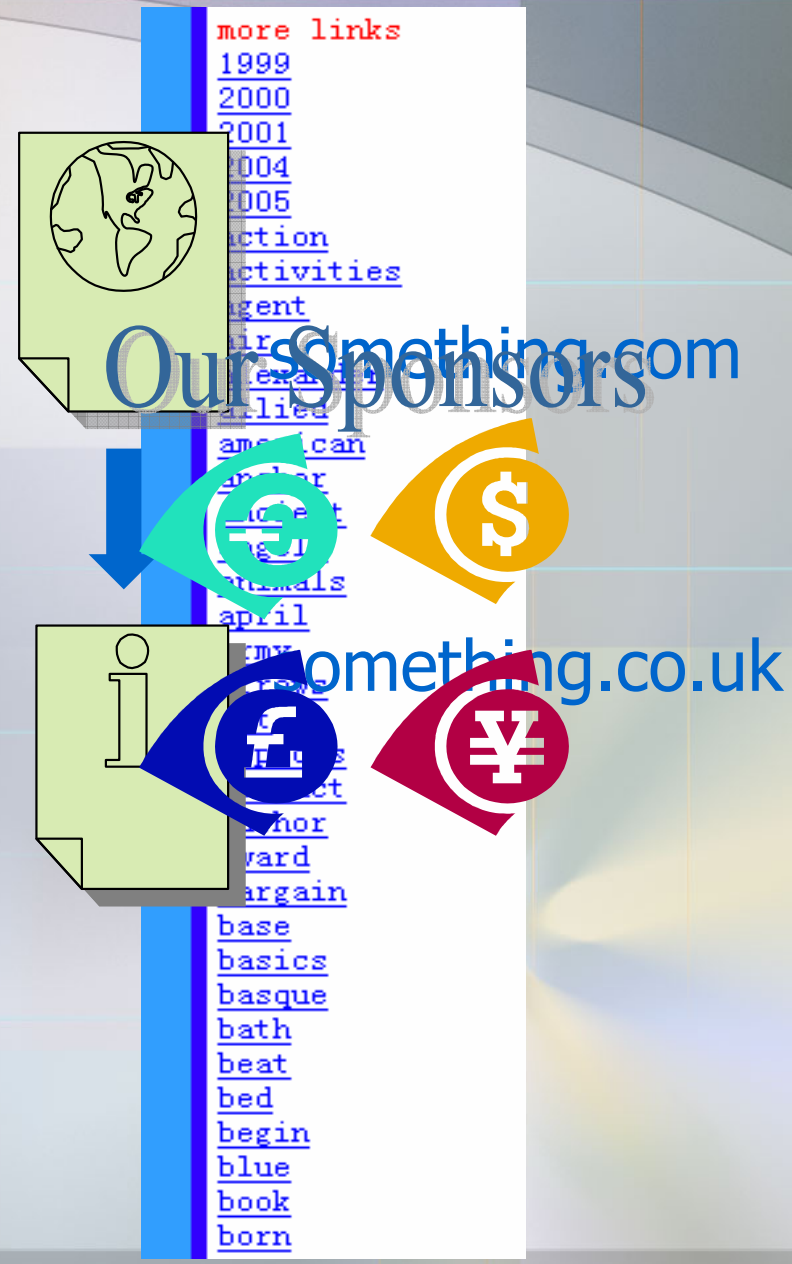
Traditional link-based ranking algorithms

(PageRank, HITS, and most variations)



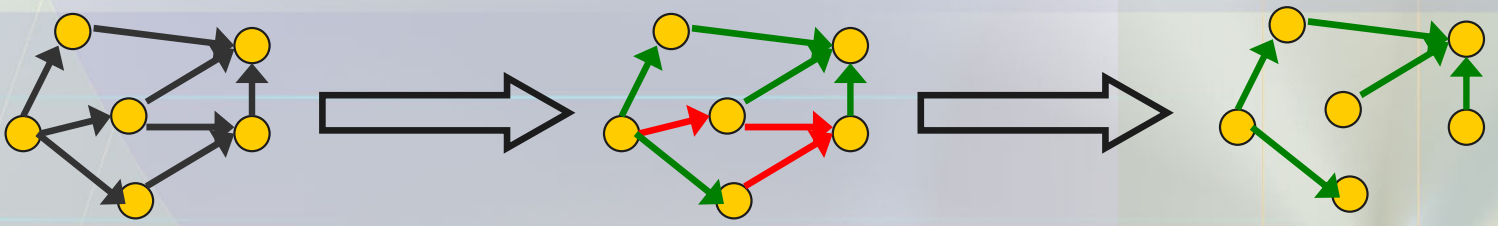
Is this always true? **No!**

- **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)**

$$A(p) = \sum_{q:q \to p} H(q) \quad \text{and} \quad H(p) = \sum_{q:p \to q} A(q)$$

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

Background (Cont.)

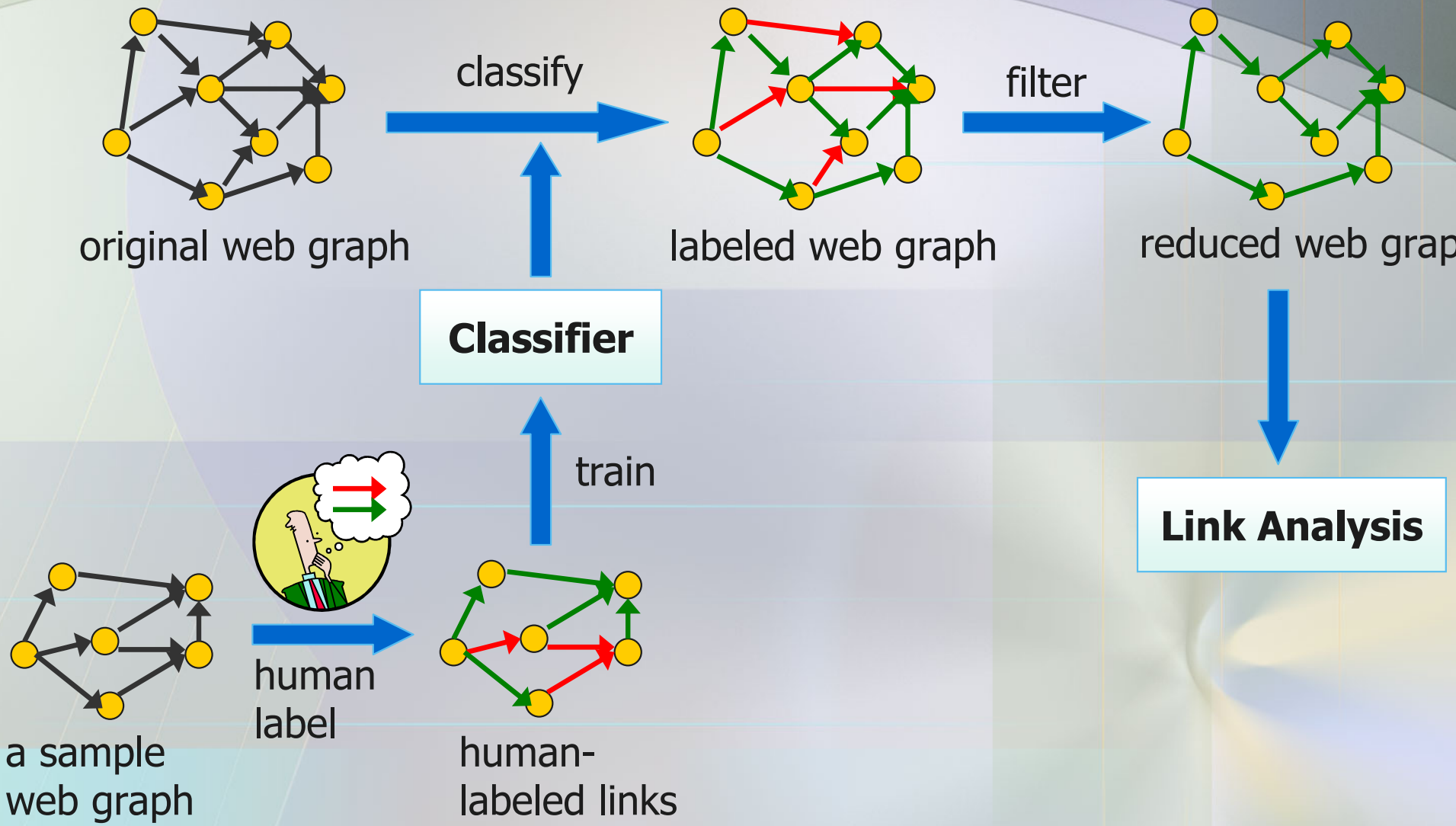
- **PageRank**
  - **Random surfer model**

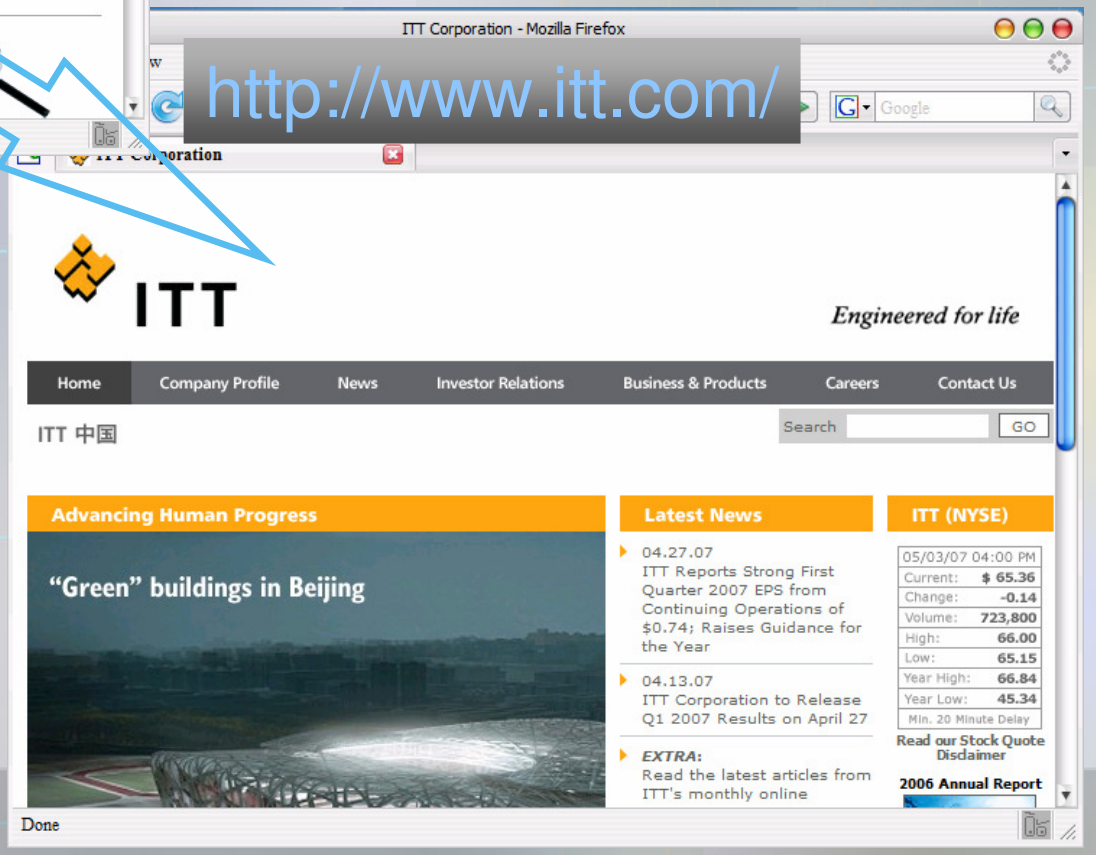
$$PR(i) = (1-d)\sum_{j:j\to i}\frac{PR(j)}{O(j)} + d\frac{1}{N}$$

Related work

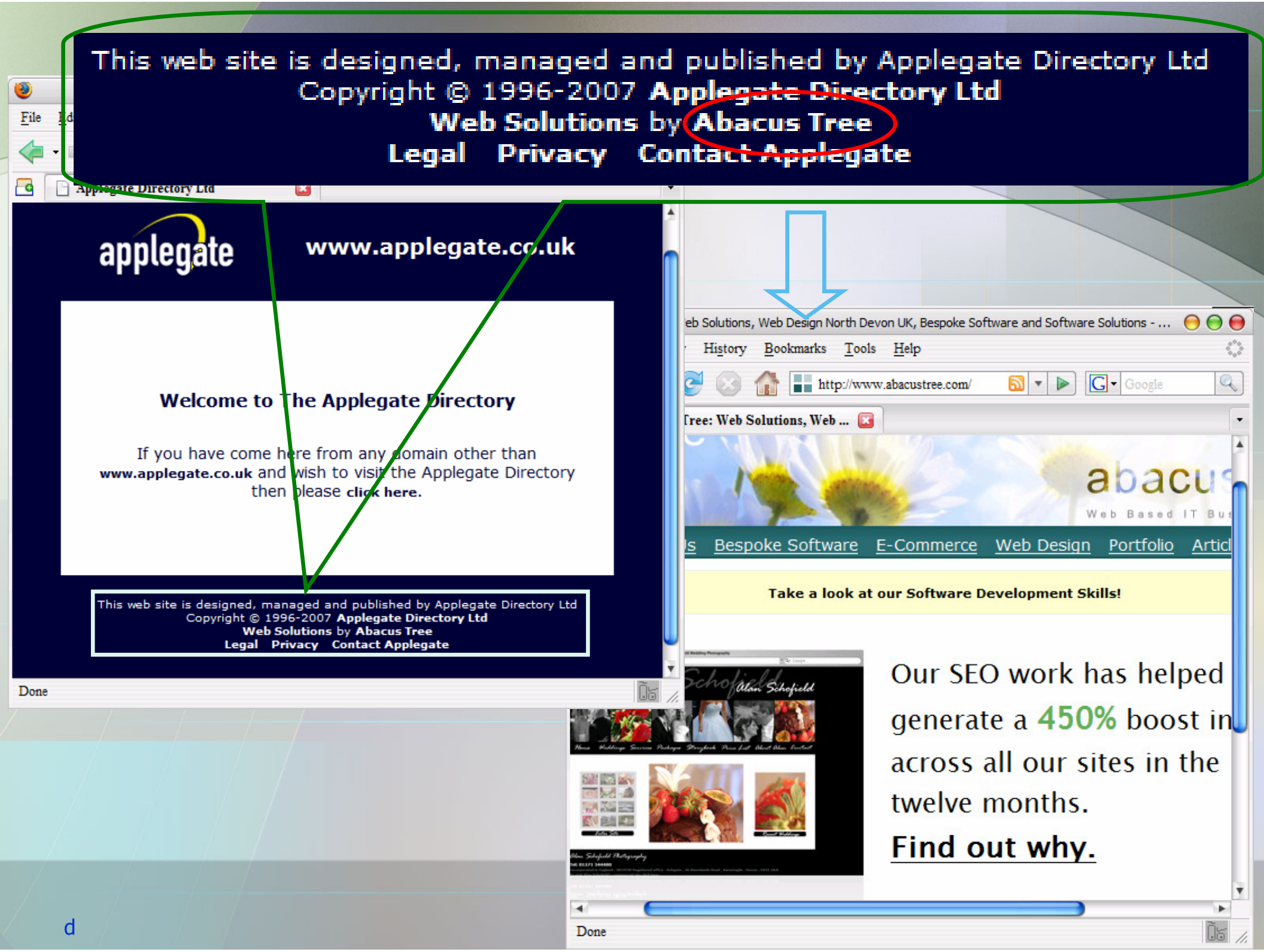
- Chakrabarti et al. (1998) improving HITS by adjusting the weights of links according to their anchor text and surrounding text.
- Davison (2000) proposed automatic recognition of nepotistic links using hand-crafted features.
- Lempel and Moran (2000) defined the tightly-knit community (TKC) effect.
- 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

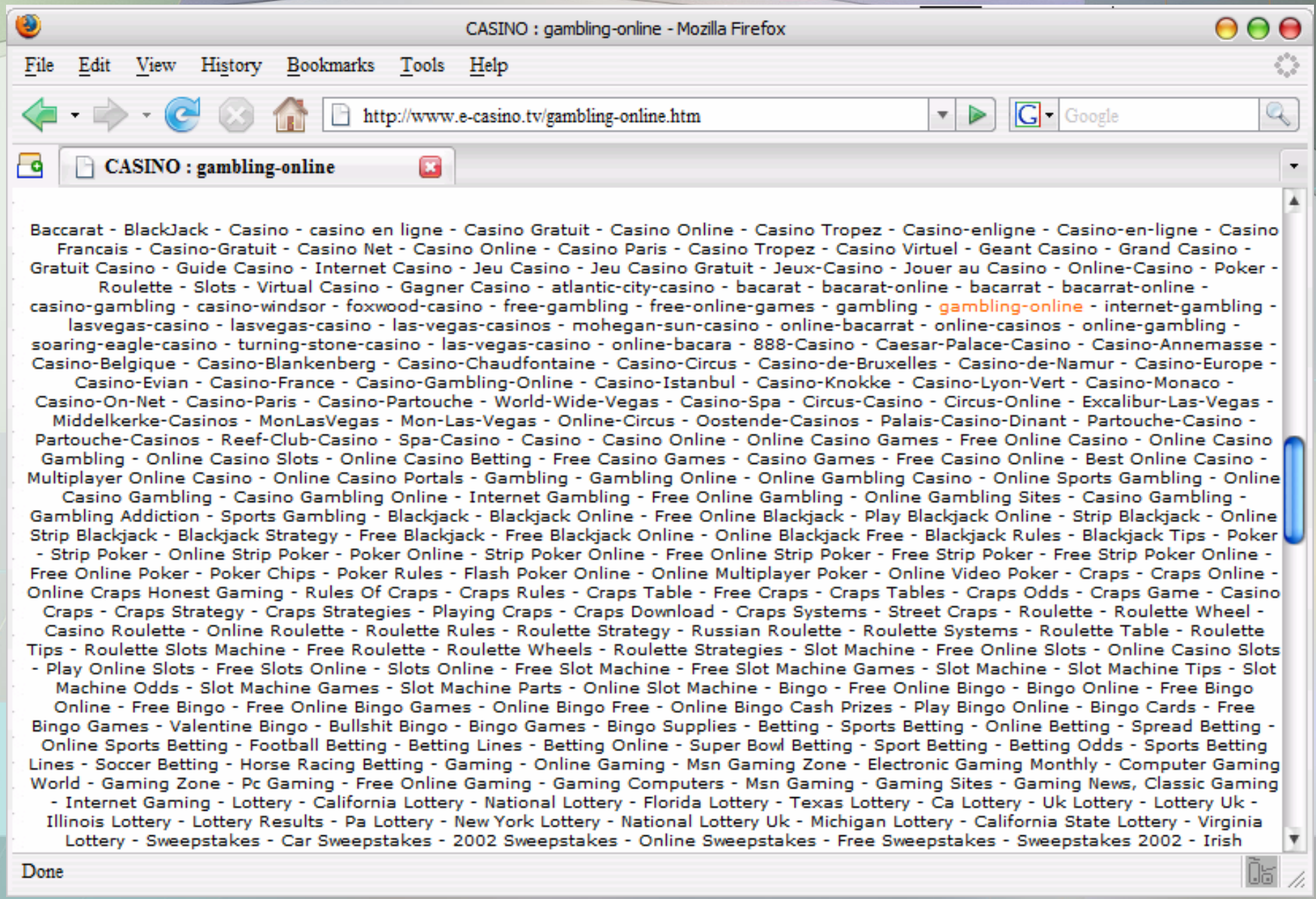

classify
filter
original web graph
labeled web graph
reduced web graph
Classifier
train
Link Analysis
human
label
a sample
web graph
human-
labeled links



A navigational link


This web site is designed, managed and published by Applegate Directory Ltd
Copyright © 1996-2007 Applegate Directory Ltd
Web Solutions by Abacus Tree
Legal Privacy Contact Applegate
applegate
www.applegate.co.uk
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.
This web site is designed, managed and published by Applegate Directory Ltd
Copyright © 1996-2007 Applegate Directory Ltd
Web Solutions by Abacus Tree
Legal Privacy Contact Applegate
Done
eb Solutions, Web Design North Devon UK, Bespoke Software and Software Solutions - ...
History Bookmarks Tools Help
http://www.abacustree.com/
Google
Tree: Web Solutions, Web ...
Bespoke Software E-Commerce Web Design Portfolio
Take a look at our Software Development Skills!
Our SEO work has helped generate a 450% boost in across all our sites in the twelve months.
Find out why.
Done

Spam links


CASINO : gambling-online - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.e-casino.tv/gambling-online.htm
Google
CASINO : gambling-online
Baccarat - BlackJack - Casino - casino en ligne - Casino Gratuit - Casino Online - Casino Tropez - Casino-enligne - Casino-en-ligne - Casino Francais - Casino-Gratuit - Casino Net - Casino Online - Casino Paris - Casino Tropez - Casino Virtuel - Geant Casino - Grand Casino - Gratuit Casino - Guide Casino - Internet Casino - Jeu Casino - Jeu Casino Gratuit - Jeux-Casino - Jouer au Casino - Online-Casino - Poker - Roulette - Slots - Virtual Casino - Gagner Casino - atlantic-city-casino - bacarat - bacarat-online - bacarrat - bacarrat-online - casino-gambling - casino-windsor - foxwood-casino - free-gambling - free-online-games - gambling - gambling-online - internet-gambling - lasvegas-casino - lasvegas-casino - las-vegas-casinos - mohegan-sun-casino - online-bacarrat - online-casinos - online-gambling - soaring-eagle-casino - turning-stone-casino - las-vegas-casino - online-bacara - 888-Casino - Caesar-Palace-Casino - Casino-Annemasse - Casino-Belgique - Casino-Blankenberg - Casino-Chaudfontaine - Casino-Circus - Casino-de-Bruxelles - Casino-de-Namur - Casino-Europe - Casino-Evian - Casino-France - Casino-Gambling-Online - Casino-Istanbul - Casino-Knokke - Casino-Lyon-Vert - Casino-Monaco - Casino-On-Net - Casino-Paris - Casino-Partouche - World-Wide-Vegas - Casino-Spa - Circus-Casino - Circus-Online - Excalibur-Las-Vegas - Middelkerke-Casinos - MonLasVegas - Mon-Las-Vegas - Online-Circus - Oostende-Casinos - Palais-Casino-Dinant - Partouche-Casino - Partouche-Casinos - Reef-Club-Casino - Spa-Casino - Casino - Casino Online - Online Casino Games - Free Online Casino - Online Casino Gambling - Online Casino Slots - Online Casino Betting - Free Casino Games - Casino Games - Free Casino Online - Best Online Casino - Multiplayer Online Casino - Online Casino Portals - Gambling - Gambling Online - Online Gambling Casino - Online Sports Gambling - Online Casino Gambling - Casino Gambling Online - Internet Gambling - Free Online Gambling - Online Gambling Sites - Casino Gambling - Gambling Addiction - Sports Gambling - Blackjack - Blackjack Online - Free Online Blackjack - Play Blackjack Online - Strip Blackjack - Online Strip Blackjack - Blackjack Strategy - Free Blackjack - Free Blackjack Online - Online Blackjack Free - Blackjack Rules - Blackjack Tips - Poker - Strip Poker - Online Strip Poker - Poker Online - Strip Poker Online - Free Online Strip Poker - Free Strip Poker - Free Strip Poker Online - Free Online Poker - Poker Chips - Poker Rules - Flash Poker Online - Online Multiplayer Poker - Online Video Poker - Craps - Craps Online - Online Craps Honest Gaming - Rules Of Craps - Craps Rules - Craps Table - Free Craps - Craps Tables - Craps Odds - Craps Game - Casino Craps - Craps Strategy - Craps Strategies - Playing Craps - Craps Download - Craps Systems - Street Craps - Roulette - Roulette Wheel - Casino Roulette - Online Roulette - Roulette Rules - Roulette Strategy - Russian Roulette - Roulette Systems - Roulette Table - Roulette Tips - Roulette Slots Machine - Free Roulette - Roulette Wheels - Roulette Strategies - Slot Machine - Free Online Slots - Online Casino Slots - Play Online Slots - Free Slots Online - Slots Online - Free Slot Machine - Free Slot Machine Games - Slot Machine - Slot Machine Tips - Slot Machine Odds - Slot Machine Games - Slot Machine Parts - Online Slot Machine - Bingo - Free Online Bingo - Bingo Online - Free Bingo Online - Free Bingo - Free Online Bingo Games - Online Bingo Free - Online Bingo Cash Prizes - Play Bingo Online - Bingo Cards - Free Bingo Games - Valentine Bingo - Bullshit Bingo - Bingo Games - Bingo Supplies - Betting - Sports Betting - Online Betting - Spread Betting - Online Sports Betting - Football Betting - Betting Lines - Betting Online - Super Bowl Betting - Sport Betting - Betting Odds - Sports Betting Lines - Soccer Betting - Horse Racing Betting - Gaming - Online Gaming - Msn Gaming Zone - Electronic Gaming Monthly - Computer Gaming World - Gaming Zone - Pc Gaming - Free Online Gaming - Gaming Computers - Msn Gaming - Gaming Sites - Gaming News, Classic Gaming - Internet Gaming - Lottery - California Lottery - National Lottery - Florida Lottery - Texas Lottery - Ca Lottery - Uk Lottery - Lottery Uk - Illinois Lottery - Lottery Results - Pa Lottery - New York Lottery - National Lottery Uk - Michigan Lottery - California State Lottery - Virginia Lottery - Sweepstakes - Car Sweepstakes - 2002 Sweepstakes - Online Sweepstakes - Free Sweepstakes - Sweepstakes 2002 - Irish
Done

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

$$Sim_{host}(x,y) = \frac{2 \times |Substr(host_x, r) \cap Substr(host_y r)|}{|Substr(host_x, r)| + |Substr(host_y r)|}$$

- **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

$$\mathrm{Sim}_{\mathrm{topic}}(x,y) = \sum_{i=1}^{n} v_{x,i} \times v_{y,i}$$

  - A topic vector $v_x = (v_{x,1}, v_{x,2}, \cdots, v_{x,n})$
  - 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

$$\mathrm{Sim}_{\mathrm{content}}(x,y) = \frac{\sum_{t \in T} (x_t \times y_t)}{\sqrt{\sum_{t \in T} x_t^2} \times \sqrt{\sum_{t \in T} y_t^2}}$$

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

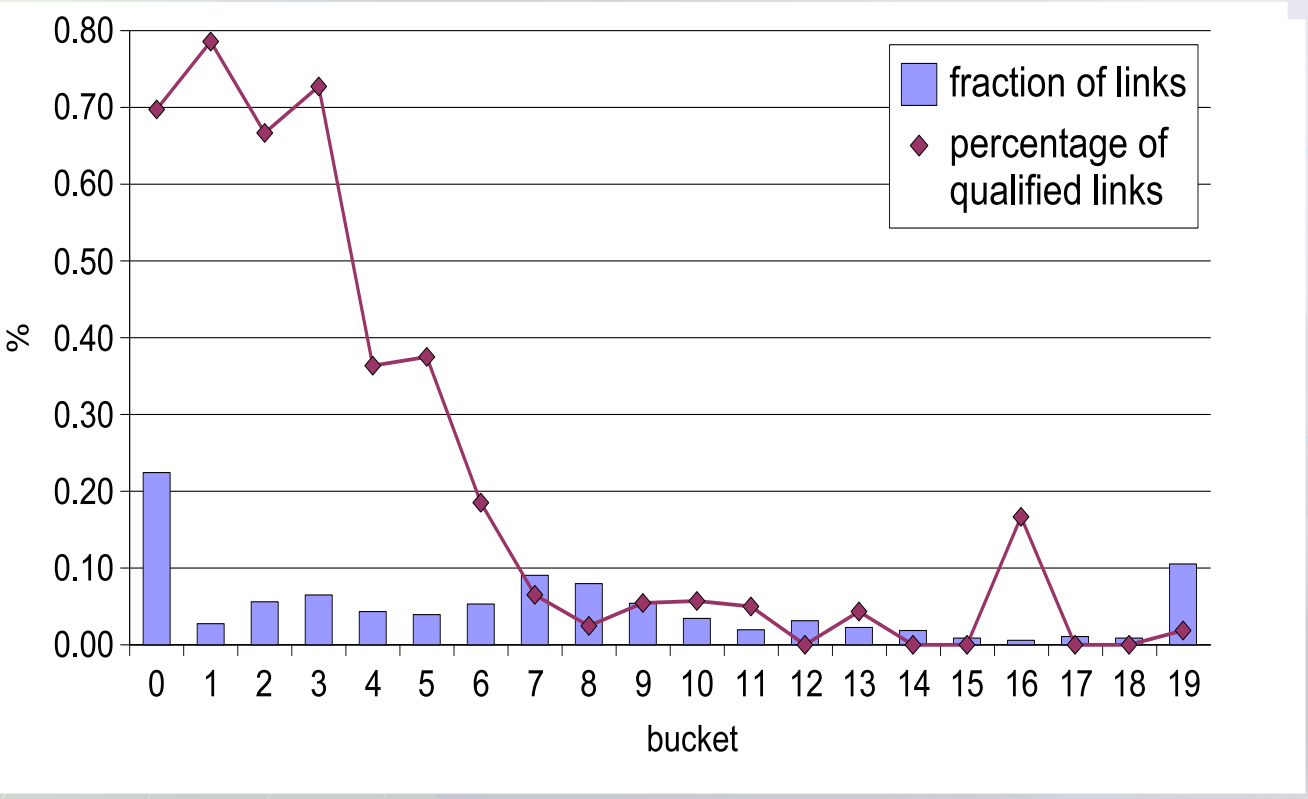
Datasets

- **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**

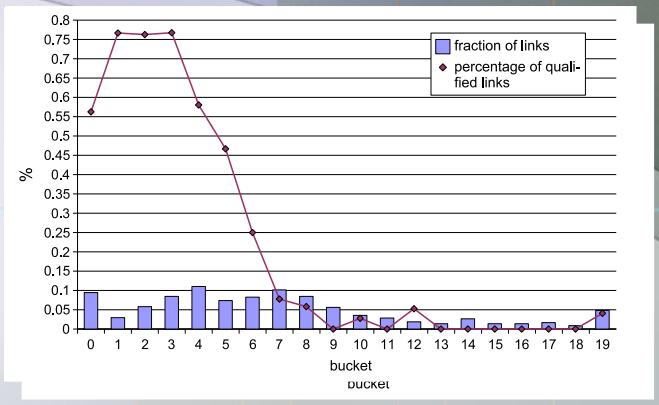
Link classification

- **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 SVM^{light}**
  - **Two fold cross validation**
  - **Average accuracy 83.8%**

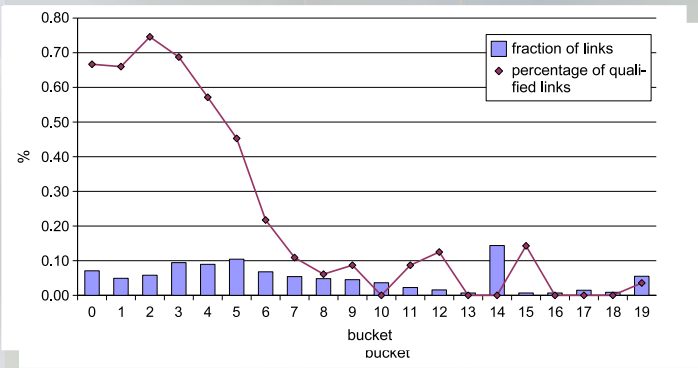
Link classification (Cont.)



Anchor text similarity



Content similarity

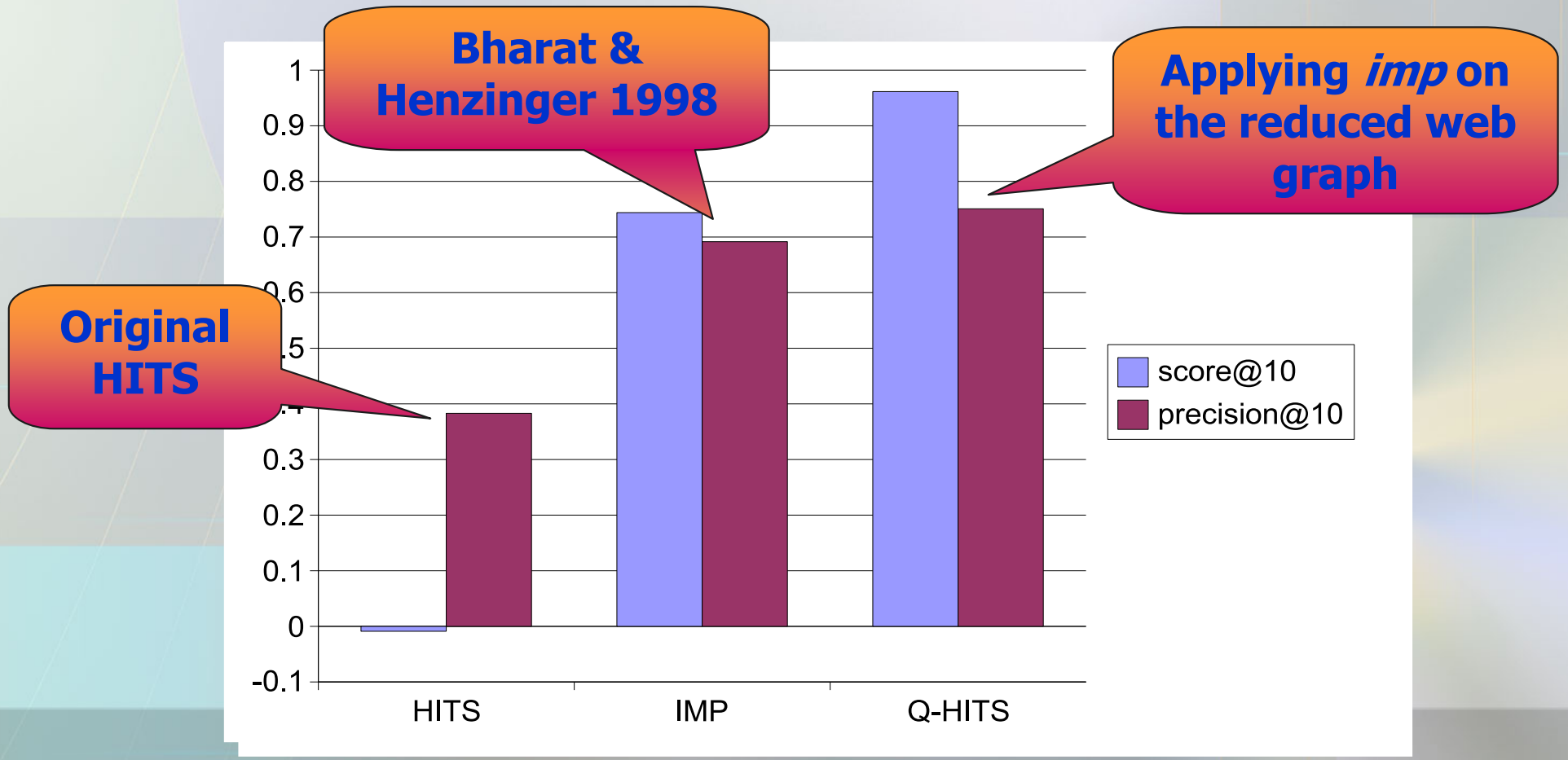


Non-anchor text similarity

Anchor text similarity was found to be the most discriminative feature.

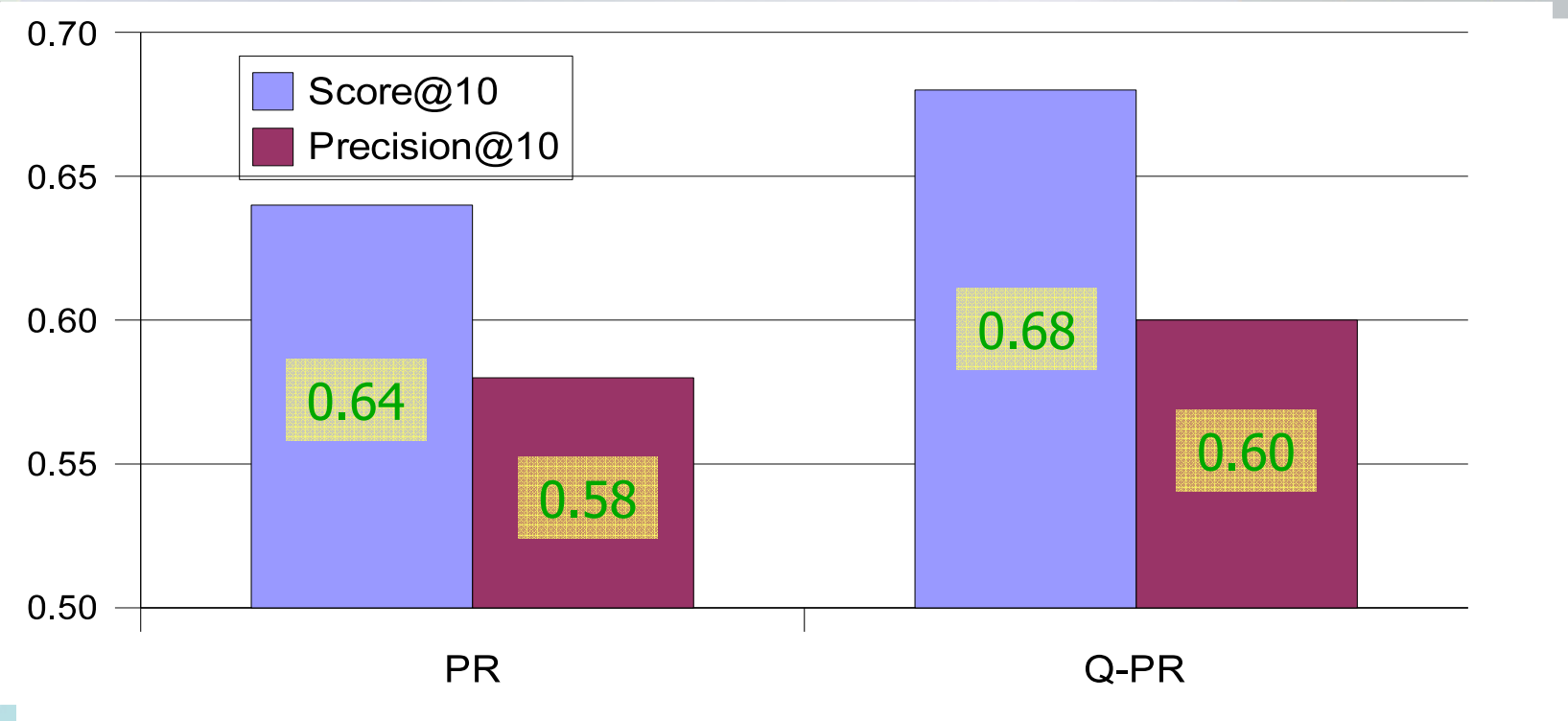
Qualified HITS

- 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.



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.



- 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/

LEHIGH UNIVERSITY